
University of Newcastle
Faculty of Science, Agriculture and Engineering
School of Electrical & Electronic Engineering

Single Channel Overlapped-Speech Detection and Separation of Spontaneous Conversations

Hasan Mohammad-Ali Kadhim

A thesis submitted to the University of Newcastle for the degree of **Doctor of Philosophy**
March – 2017



I present my PhD thesis to:

- My country **IRAQ**
- **My Family** in Iraq and the UK

Table of Content

Table of Content.....	i
Abstract	v
Acknowledgment	vii
List of Figures	ix
List of Tables	xi
List of Symbols, Abbreviations and Acronyms	xiii
List of Publications	xvii
List of Hibernating	xix
Chapter 1. Introduction	1
1.1 Structure of the Thesis.....	1
1.2 Speech versus Audio	2
1.3 Pitch of Speech Signal	4
1.4 Spontaneous Conversation, Dialog Speech and Mixture Speech.....	6
1.5 Overlapped-Speech Detection, Speaker Diarization and Speech Separation	8
1.6 Samples, Window-Frame and Hopping period.....	8
1.7 Supervised, Semi-Supervised and Unsupervised Machine Learning.....	10
1.8 Blind Speech Separation versus Informed Speech Separation.....	11
1.9 Overall-System	12
1.10 Subjective Test versus Objective Test.....	15
1.11 Masking.....	17
1.12 Objectives and Aims of the Research	19
1.13 Contributions	20
Chapter 2. Literature Reviews.....	23
2.1 Introduction	23
2.2 Literature Review of Overlapped-Speech Detection	24
2.3 Literature Review of Speech Separation by Non-negative Matrix Factorization	31
2.4 Literature Review of Informed Speech Separation	35
2.4.1 <i>Video-Assisted Source Separation</i>	35
2.4.2 <i>Spatial Audio Object Coding</i>	36
2.4.3 <i>Reverberant Models for Source Separation</i>	37
2.4.4 <i>Score-Informed Source Separation</i>	38

2.4.5	<i>Language-Informed Speech Separation</i>	38
2.4.6	<i>User-Guided Source Separation</i>	39
2.4.7	<i>Dictionary-Based Methods</i>	39
2.4.8	<i>NMF-Based Informed Speech Separation</i>	40
Chapter 3.	Overlapped-Speech Detection based-on Stochastic Properties	43
3.1	Introduction	43
3.2	Functional Block Diagram and Illustrative Waveforms	44
3.3	An Algorithm of Overlapped-Speech Detection	45
3.3.1	<i>Framing and Overlapping-Window of the Input Signal</i>	46
3.3.2	<i>Extraction of Audio Features by RASTA-PLPC</i>	46
3.3.3	<i>k-means Clustering of the Features</i>	52
3.3.4	<i>Groups and Statistical Variances</i>	58
3.3.5	<i>Optimizing the Groups</i>	60
3.3.6	<i>Re-clustering</i>	64
3.3.7	<i>Hierarchical Clustering Scenarios</i>	65
3.4	Experiments	67
3.5	Result and Test	70
3.6	Comparison	76
3.7	Summary	80
Chapter 4.	Blind Speech Separation by Filter-Bank, Non-negative Matrix Factorization and Speaker Clustering	83
4.1	Introduction	83
4.2	Source Separation	84
4.3	Functional Block Diagrams and Waveforms	86
4.3.1	<i>Preparation of the Required Resources</i>	88
4.3.2	<i>Filter-Bank Analysis Technique</i>	89
4.3.3	<i>Non-negative Matrix Factorization NMF</i>	91
4.3.4	<i>Speaker Clustering</i>	94
4.4	Experiments	95
4.5	Result and Test	100
4.6	Comparison	113
4.7	Summary	114
Chapter 5.	Informed Speech Separation by Semi-Supervised Non-negative Matrix Factorization	115
5.1	Introduction	115
5.2	Functional Block Diagrams and Waveforms	116
5.3	Informed Speech Separation Procedure	117
5.3.1	<i>Preparation of the Required Resources</i>	118

5.3.2	<i>Training the Virtual Speech Signals</i>	119
5.3.3	<i>The Virtual Assists the Real Speech Signals</i>	121
5.3.4	<i>Soft and Binary Masking</i>	122
5.3.5	<i>Exploiting both Masks</i>	123
5.4	Experiments	125
5.5	Results and Tests	126
5.6	Comparison	142
5.7	Summery	144
Chapter 6. Notes, Conclusions and Future Works		145
6.1	Notes and Conclusions	145
6.2	Future Works	146
Appendix A. Historical Overviews		149
A.1	Historical Overview of Filter-Bank	149
A.2	Historical Overview of k-means	149
A.3	Historical Overview of Overlapped-Speech Detection	150
A.4	Historical Overview of Speech Separation	150
A.5	Historical Overview of NMF and NMF-based Speech Separation	151
Appendix B. Software		153
Appendix C. k-means Flowchart		155
Appendix D. NMF Procedure		157
References		159

Abstract

In the thesis, spontaneous conversation containing both speech mixture and speech dialogue is considered. The speech mixture refers to speakers speaking simultaneously (i.e. the overlapped-speech). The speech dialogue refers to only one speaker is actively speaking and the other is silent. That Input conversation is firstly processed by the overlapped-speech detection. Two output signals are then segregated into dialogue and mixture formats. The dialogue is processed by speaker diarization. Its outputs are the individual speech of each speaker. The mixture is processed by speech separation. Its outputs are independent separated speech signals of the speaker. When the separation input contains only the mixture, blind speech separation approach is used. When the separation is assisted by the outputs of the speaker diarization, it is informed speech separation. The research presents novel: overlapped-speech detection algorithm, and two speech separation algorithms.

The proposed overlapped-speech detection is an algorithm to estimate the switching instants of the input. Optimization loop is adapted to adopt the best capsulated audio features and to avoid the worst. The optimization depends on principles of the pattern recognition, and k-means clustering. For of 300 simulated conversations, averages of: False-Alarm Error is 1.9%, Missed-Speech Error is 0.4%, and Overlap-Speaker Error is 1%. Approximately, these errors equal the errors of best recent reliable speaker diarization corpuses.

The proposed blind speech separation algorithm consists of four sequential techniques: filter-bank analysis, Non-negative Matrix Factorization (NMF), speaker clustering and filter-bank synthesis. Instead of the required speaker segmentation, effective standard framing is contributed. Average obtained objective tests (SAR, SDR and SIR) of 51 simulated conversations are: 5.06dB, 4.87dB and 12.47dB respectively.

For the proposed informed speech separation algorithm, outputs of the speaker diarization are a generated-database. The database associated the speech separation by creating virtual targeted-speech and mixture. The contributed virtual signals are trained to facilitate the separation by homogenising them with the NMF-matrix elements of the real mixture. Contributed masking optimized the resulting speech. Average obtained SAR, SDR and SIR of 341 simulated conversations are 9.55dB, 1.12dB, and 2.97dB respectively.

Per the objective tests of the two speech separation algorithms, they are in the mid-range of the well-known NMF-based audio and speech separation methods.

Acknowledgment

My thanks to “Al-Mustansiriyah University”, when the university candidate my name for its study-leave program. My thanks to “Newcastle University” which existed my dream to enroll with their PhD-program.

I appreciate the unlimited supports of “my family in Iraq and in the UK”. They were easing the hard in my PhD-program.

My respects to my supervisors: Dr Lok and Prof Satnam, for their advices during the PhD-program (*When you teach me something, you have owned me a slave / Arabic heritage*). I have liked the DSP recently; I love the DSP now.

I will not forget the lovely “staff of the Electrical and Electronic Engineering School, Newcastle University”.

Many thanks to “The UK” and the wonderful city “Newcastle-Upon-Tyne” for their kind hospitality.

I would like to mention the assistance of “my friends in Iraq and the UK”.

The Black, the White and the Grey were wonderful colours for my thesis.

List of Figures

Figure 1.1 Typical sketch of statistical distribution of the harmonics for long time of speech.	5
Figure 1.2 Typical STFT Time-Frequency Spectrogram for 2 second of speech.	6
Figure 1.3 Arbitrary spontaneous conversation.	7
Figure 1.4 Typical machine learning DSP systems.	11
Figure 1.5 Chapter 4 overall system.	12
Figure 1.6 Chapter 5 overall system.	13
Figure 1.7 Typical sketches of the overlapped-speech detection process.	14
Figure 1.8 Typical sketches of the speaker diarization process.	14
Figure 1.9 Typical sketches of the speech separation process.	14
Figure 1.10 Typical speech-DSP output signals for the spontaneous conversation.	14
Figure 1.11 Typical sketches of Aims and Objectives of the research.	21
Figure 3.1 Typical sketches of input and outputs of Chapter 3 algorithm.	44
Figure 3.2 General block diagram of the overlapped speech detection process.	45
Figure 3.3 Flow chart of the first Hermansky version of Perceptual Linear Prediction.	49
Figure 3.4 The normalized-scale curves of the Critical-Bands analysis.	49
Figure 3.5 Flow chart of the simplified version of RelAtive-SpecTrAl (RASTA) filter.	51
Figure 3.6 Flow chart of the RASTA-Perceptual Linear Predictive Coefficients (RASTA-PLP).	52
Figure 3.7 Audio features extraction, and Initial crude clustering.	54
Figure 3.8 Specimens of three PDFs.	56
Figure 3.9 The Grouping concept.	57
Figure 3.10 Sketches illustrate the effect of the distances of the centroids of two PDFs.	62
Figure 3.11 Sketches illustrate the recognition between two distribution patterns.	62
Figure 3.12 Per-Unit Goodness of the recognition between the patterns of the variances of the audio features of spontaneous conversation between F&M.	64
Figure 3.13 Typical scheme of the two Hierarchical Clustering Scenarios.	66
Figure 3.14 Flowchart of Chapter 3 algorithm.	68
Figure 3.15 The waveforms of the implementation of Chapter 3 algorithm.	71
Figure 3.16 The percentage average DER.	75
Figure 3.17 The implementation when the speech has a long period of silence during the mixture speech.	77
Figure 3.18 Missed-Speech Error Rate (E_{MISS}) comparison between the research algorithm with the standard speaker diarization corpuses	79
Figure 4.1 Chapter 4 overall system.	86
Figure 4.2 Arbitrary spontaneous conversation.	87
Figure 4.3 Functional block diagram of this chapter algorithm.	88
Figure 4.4 Block diagram of this chapter algorithm.	91
Figure 4.5 Flowchart of Chapter 4 algorithm.	96
Figure 4.6 Specimen of the filter-bank analysis.	98

Figure 4.7 Specimen of the NMF speech separation.....	99
Figure 4.8 Speech signal waveforms of the Female (TIMIT).....	102
Figure 4.9 Spectrograms of the Female (TIMIT), for the 1 kHz range..	103
Figure 4.10 7s-500Hz T-F Spectrogram of the tested Female.	104
Figure 4.11 Speech signal waveforms of the Male (TIMIT).	105
Figure 4.12 Spectrograms of the Male (TIMIT), for the 1kHz range.....	106
Figure 4.13 7s-500Hz T-F Spectrogram of the tested Male.	107
Figure 4.14 Graphic bars represent the minimum values of the objective tests.	108
Figure 4.15 Graphic bars represent the maximum values of the objective tests.	109
Figure 4.16 Graphic bars represent the average values of the objective tests..	110
Figure 4.17 The average values of objective tests.....	111
Figure 4.18 Graphic bars represent the average values of the SAR, the SDR and the SIR (dB)...	112
Figure 4.19 One-second waveforms to compare between the binary masking and the soft masking..	113
Figure 5.1 Chapter 5 overall system.	116
Figure 5.2 Chapter 5 arbitrary spontaneous conversation.....	117
Figure 5.3 Functional block diagram of chapter 5 system.....	118
Figure 5.4 Optimization functional block diagram to improve the fluctuating of the objective tests.....	124
Figure 5.5 Waveforms of mixture, targeted and recovered speech of the 1 st speaker.	127
Figure 5.6 Spectrograms of mixture, targeted and recovered speech of the 1 st speaker.....	128
Figure 5.7 Waveforms of mixture, targeted and recovered speech of the 2 nd speaker.....	129
Figure 5.8 Spectrograms of mixture, targeted and recovered speech of the 2 nd speaker.....	130
Figure 5.9 SAR tests (dB) of the Chapter 5 algorithm for all the 341 conversations.....	132
Figure 5.10 SDR tests (dB) of Chapter 5 algorithm for all the 341 conversations.	134
Figure 5.11 SIR tests (dB) of Chapter 5 algorithm for all the 341 conversations.	136
Figure 5.12 The optimized objective tests of Chapter 5 algorithm for 341 conversations.	138
Figure 5.13 The optimized objective tests of Chapter 5 algorithm for 341 conversations.....	139
Figure 5.14 Variances of data of the SAR, SDR and SIR 341 tests.....	140
Figure 5.15 Maximum, average, minimum and variance values of the tests.....	141
Figure 5.16 compression with recent tests of well-known articles.	143

List of Tables



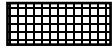
Table 3.1 V-array arrangement..	60
Table 3.2 The percentage average DER without the algorithm.	74
Table 3.3 The percentage average DER using the algorithm.	74
Table 3.4 comparison between the research algorithm with recent articles.	78
Table 4.1 Minimum average values of the objective tests..	108
Table 4.2 Maximum average values of the objective tests.	109
Table 4.3 Average values of the objective tests.	110
Table 5.1 SAR objective tests (dB) of Chapter 5 algorithm.	131
Table 5.2 SDR objective tests (dB) of the Chapter 5 algorithm.	133
Table 5.3 SIR objective tests (dB) of the Chapter 5 algorithm.	135
Table 5.4 The 2 nd speaker objective tests of Chapter 5 algorithm.	137
Table 5.5 Comparison with recent well-known articles.	142

List of Symbols, Abbreviations and Acronyms

A_v	Average value.
BIC	Bayesian Information Criterion.
bit	binary digit.
BSS	Blind Speech (Source) Separation.
BW	Band-width.
CASA	Computational Auditory Scene Analysis.
C	number of matrix [H] columns.
dB	decibel.
D_{cent}	Distance between two centroids.
DER	Diarization Error Rate.
DFT	Discrete Fourier Transform.
DNN	Deep Neural Network.
d_f	Euclidian distance from F-parameters to the reference-parameters.
d_m	Euclidian distance from M-parameters to the reference-parameters.
D-vector	Decision-vector, a vector contains the k-means clustering of variances.
e_{inter}	Interference error.
e_{noise}	Noise error.
e_{artif}	Artifact error.
E_{OVL}	(dialogue speech which is identified as mixture speech) plus (mixture speech which is identified as dialogue speech).
E_{FA}	False-Alarm-Rate (FAR). Error of a dialogue inside the mixture.
E_{MISS}	error is caused by a mixture inside the dialogue.
ERB	Equivalent Rectangular Bandwidth frequency scale.
$dur(.)$	duration of.
$e(n)$	discrete error.
$floor(.)$	down-rounding floor function.
F_r	Real Male targeted-speech.
F	Female alone (dialogue) speech (frequency domain).
FM	Female and Male simultaneously (mixture) speech (frequency domain).
FM_r	Real Mixture F with M speech simultaneously.
FM_v	Virtual F and M simultaneously, mixture speech.
F_v	Virtual Female targeted-speech.
f_s	Sampling rate (frequency) by Hz.
F_v	Virtual female speech.
F_{Ed}	Euclidian distance from F to the reference.
G ₁ to G ₃₂	0.1-s to 3.2-s group of features.

GitHub	open source speaker diarization tool-box (by the GitHub inst.).
GMM	Gaussian Mixture Modelling.
<i>Goodness</i>	Relative Goodness of pattern recognition.
$[H]$	Activation-Weights matrix.
$[H_v]$	Virtual Activation-Weights matrix.
HMM	Hidden Markov Modelling.
Hz	Hertz, SI unit of the frequency.
ICA	Independent Component Analysis.
IDCT	Inverse Discrete Fourier Transform.
i/p	input signal.
<i>inv(.)</i>	Matrix inverse function/MATLAB.
ISS	Informed Speech (Source) Separation.
ISTFT	Inverse Short Time Fourier transform.
k-means	k-means parallel clustering.
k-means++	k-means plus-plus clustering.
<i>K</i> -vector	a vector contains the k-means clustering of features.
LFCC	Linear-Frequency Cepstral Coefficients.
LPC	Linear Predictive Coding.
M	Male alone (dialogue) speech (frequency domain).
MFCC	Mel-Frequency Cepstral Coefficients.
ML	Machine Learning.
M_r	Real Male targeted-speech.
MSE	Mean Square of the Error.
M_v	Virtual Male targeted-speech.
m_v	Virtual male speech.
M_{Ed}	Euclidian distance from M to the reference.
N_c	Number of coefficients (features) per frame.
N_f	Number of frames per conversation.
N_t	Number of total samples per conversation.
N_h	Number of samples each hop.
N_{ref}	Number of speaker speaking in reference segments.
$N_{correct}$	Number of speaker speaking in correct segments.
N_{sys}	Number of speaker speaking in segments.
$N_{correct}$	Number of speakers those speaking in segment.
N_{ts}	total number of speech segments (Hierarchical Scenario).
N_{ss}	starting number of speech segments (Hierarchical Scenario).
N_{es}	ending number of speech segments (Hierarchical Scenario).
N_{spi}	Number of speakers per the i^{th} segment.

N_s	Number of the processed segment.
N_{sb}	Number of filter-bank sub-bands.
NMF	Non-Negative Matrix Factorization.
N_w	Number of samples per overlapping-windowed frames.
OOP	Object Oriented Programming.
o/p	output signal.
RAPT	<u>Robust Algorithm for Pitch Tracking.</u>
PCA	Principal Component Analysis.
PCM	Pulse Code Modulation.
PDA	Pitch Detection Algorithm.
PDF	Probability Density Function.
PLPC	Perceptual Linear Prediction Coefficients.
PNCC	Power-Normalised Cepstral Coefficients.
PR	Pattern Recognition.
PU	Per Unit.
pyAudioAnalysis	Open-Source Python Library for Audio Signal Analysis.
$pinv(.)$	The Moore-Penrose inverse (pseudo-inverse) of symbolic matrix, MATLAB function.
r	number of matrix [W] rows.
RASTA	RelAtive-SpecTrAl filter.
RHS, LHS	Right-Hand Side, Left-Hand Side.
RT02 to RT09	Rich-Transcription series (from RT02 to RT09).
s	second, SI unit of the time.
[S]	Spectrogram Time-Frequency domain matrix.
ss	number of matrix [W] columns = number of matrix [H] rows.
s_{target}	Reference signal (speech separation)/ targeted-speech.
s.d.p.d.	State-Dependent Probability Distribution.
SAR	energy Source to Artifacts Ratio.
SDR	energy Source to Distortion Ratio.
SIR	energy Source to Interferences Ratio.
SNR	energy Source to Noise Ratio.
SAR-FB	} SAR, SDR and SIR for the female targeted signal, using binary masks.
SDR-FB	
SIR-FB	
SAR-FS	} SAR, SDR and SIR for the Female targeted signal, using soft masks.
SDR-FS	
SIR-FS	

SAR-MB	}	SAR, SDR and SIR for the male targeted signal, using binary masks.
SDR-MB		
SIR-MB		
SAR-MS	}	SAR, SDR and SIR for the male targeted signal, using Mask masks.
SDR-MS		
SIR-MS		
SNMF		Sparse Non-Negative Matrix Factorization.
SMCR		Self-Modeling Curve Resolution.
STFT		Short Time Fourier Transform.
τ		time width of group-period.
T_s		duration of the mixture speech segment.
T_f		duration of the overlapping-window speech frame.
T_h		duration of the hopping.
TIMIT		Audio and speech library of MIT institute.
TVAD		Theorizing Visual Art and Design simulation software.
V-vector		a vector contains variances.
V_1 to V_{32} etc.		Variance 0.1-s to 3.2-s of group of features.
$[W]$		Spectral-Basis matrix.
$[W_v]$		Virtual spectral-basis matrix.
$x(n)$		time-domain discrete signal.
$X(k)$		frequency-domain discrete signal.
		Speech segment (section) of female alone (dialogue).
		Speech segment (section) of male alone (dialogue).
		Speech segment (section) of female and male simultaneously (mixture).
$[\varphi_{xr}]$		Phase-angle matrix for the NMF
\mathbb{R}		Real-number values.

List of Publications

1. H. A. Kadhim, L. Woo, and S. Dlay, " Overlapped-speech detection and blind speech separation of spontaneous conversation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, submitted, 2017.
2. H. A. Kadhim, L. Woo, and S. Dlay, " Overlapped-speech detection and informed speech separation of spontaneous conversation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, submitted, 2017.
3. H. A. Kadhim, L. Woo, and S. Dlay, "Speech separation of spontaneous conversation by filter-bank, NMF and speaker clustering," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, submitted, 2017.
4. H. A. Kadhim, L. Woo, and S. Dlay, "Informed speech separation of spontaneous conversation by semi-supervised NMF," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, submitted, 2017.
5. H. A. Kadhim, L. Woo, and S. Dlay, "Stochastic overlapped-speech detection of spontaneous conversation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, submitted, 2017.
6. H. A. Kadhim, L. Woo, and S. Dlay, "Statistical speaker diarization using dependent combination of extracted features," in *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, 2015, pp. 291-296.
7. H. A. Kadhim, L. Woo, and S. Dlay, "Statistical speech segregation using the developed k-means of audio feature," presented at the *IEEE 3rd International Conference on Image Information Processing (ICIIP)*, 2015.
8. H. A. Kadhim, L. Woo, and S. Dlay, "Novel algorithm for speech segregation by optimized k-means of statistical properties of clustered features," in *2015 IEEE 2nd International Conference on Progress in Informatics and Computing (PIC)*, 2015, pp. 286-291.
9. H. A. Kadhim, L. Woo, and S. Dlay, 2016. Speaker diarization by dependent combination of audio features. *International Journal of Simulation--Systems, Science & Technology IJtauT*, 16(1).

List of Hibernating

- *Table of Content, List of Figures* and *List of Tables* hibernate onto their items.
- The citations hibernate onto their references.
- The following are hibernated from their cross-references: Figures, Tables and Equations.
- The following are hibernated: main titles (e.g. the Lists and the Chapters), sub-titles (e.g. the sub-title in the chapters and the sub-titles in the appendices) and sub-sub-titles.
- Websites of: some references and related researchers, speaker diarization corpuses, official websites of the experiments' software (e.g. MATLAB) and main toolboxes which are used for invoking the required codes. The websites' hibernating is underlined.

Chapter 1. Introduction

1.1 Structure of the Thesis

In addition to the tables, the lists and appendices, the thesis contains the following: **Chapter 1 “Introduction”** prefaces the thesis to readers. Several sub-titles inside Chapter 1 could be included inside other chapters, but I prefer to state them inside Chapter 1. **Chapter 2 “Literature Reviews”** contains the most important and newest literature reviews of the terms: Overlapped-speech detection, NMF-based blind speech separation and NMF-based informed speech separation. Chapter 3, Chapter 4 and Chapter 5 are kernel of the theses. The three main research contributions are detailed in those chapters. **Chapter 3 “Overlapped-Speech Detection based-on Stochastic Properties”** details novel algorithm to detects, and then segregates the spontaneous conversations into: mixture and dialogue speech formats. **Chapter 4 “Blind Speech Separation by Filter-Bank, Non-negative Matrix Factorization and Speaker Clustering”** covers novel algorithm for single channel blind speech separation of the mixture format of the Chapter 3 outputs. **Chapter 5 “Informed Speech Separation by Semi-Supervised Non-negative Matrix Factorization”** contains novel algorithm for informed speech separation of the mixture format of the Chapter 3 outputs. The dialogue format of Chapter 3 outputs should be an input signal of a speaker diarization process. The outputs of the speaker diarization are the database-like for the Chapter 5 algorithm. In addition to the five chapters, **Chapter 6 “Notes, Conclusions and Future Works”** are for Chapter 3, Chapter 4 and Chapter 5.

There are links between Chapter 3, Chapter 4 and Chapter 5 materials. The observation signal of the research thesis is 2-speaker saponaceous conversation. This signal is the input of Chapter 3 algorithm. The first output signal of the algorithm is a mixture speech, which is the input signal of the speech separation algorithms in Chapter 4 and Chapter 5. The second output of Chapter 3 algorithm is a dialogue speech, which is the input of a speaker diarization. The speaker diarization is invoked from existing toolbox. The thesis does not detail the speaker diarization. Outputs of the speaker diarization support Chapter 5 algorithm, but do not support Chapter 4 algorithm. Hence, Chapter 5 contains informed speech separation and Chapter 4 contains blind speech separation. For appendices (A to D) have been added to the thesis in order to support the material of the six chapters of the thesis.

1.2 Speech versus Audio

According to the written human kind history, speech is the fastest, the simplest, the efficient and the most emotional communicating tool. Speech has been having an excellent ability to describe various human feelings in the human life. In signal processing, although the speech is part of the audio and they have common algorithms, but speech signal processing has different approaches compared with audio signal processing. On the other hand, the speech and the audio signal processing have their unique algorithms and approaches compared with other digital signal processing (e.g. image signal processing). In time domain, speech and/or audio signals could be described as a sequence of discrete encoded samples. Sampling rate (frequency) of speech is 8000 sample/second for ordinary DSP and telecommunication, and 16000 sample/s for highly defined DSP and telecommunication. Practically, those rates had been chosen according to laboratory experiments. Theoretically, they are chosen according to the Nyquist-Shannon sampling theorem. The condition of this theorem is: to convert any analogue signal to its equivalent digital format, the sampling rate must be more than twice its bandwidth. This condition is proved to reduce the distortion by avoiding the aliasing in the frequency domain. The aliasing is an overlapping of each sub-band harmonic with its two adjacent: left hand side and right hand side sub-bands. So, the taken in account bandwidth of the original speech signal is 4 kHz for ordinary speech signal and 8 kHz for highly defined speech signal [1, 2]. Audio signal includes: speech and any other heard (audible) signals such as music, sound of machine, sound of animals ... etc. Human ear can hear an acoustical signal up to 22 kHz. The maximum sampling rate is 44000 sample/second. That rate is converting analog audio signal to its equivalent highly-defined discrete format [1, 2].

The speech is a physical phenomenon (mechanical movements) which is produced by the human voice system [1, 2]. The system consists of biological components inside and outside human body. Speech production begins by diaphragm vibrations and finishes by lips movements, so speech characteristics of speakers have variety and non-similar. Due to those variety and the non-similarity, it's difficult to recognize the audio DSP parameters of different speech and different speakers. Speech is a sequence of letters, words, compounds of words and sentences, so enough databases of these sequences assist the recognition job among the speech of one and different speakers. Recognition job among speakers is more challenge than among speech. Speech and speaker recognition have a lot of limitations because most of speech energy occupies a narrow band of the frequency domain (less than 2 kHz) [3].

The scientific dealing with audio has less challenge than with speech, because the typical audio bandwidth is 3 to 5 times wider than speech bandwidth. Also, energy of audio signal occupies different sub-bands of its frequency domain spectrum, but energy of the speech, almost occupies the lower frequency sub-bands.

According to nature of the human voice system and dimensions of its components, best interval of speech processing is 8 to 20 ms (10^{-3} second). The interval formulates a short frame of speech signal in the Short Time Fourier transform STFT analysis. For that interval, the sub-band resolution of the processed signal is 50 to 125 Hz. This resolution is not adequate to process the speech-DSP algorithms efficiently. The overlapping-window is a technique to increase the resolution twice by increasing the frame interval to (16 to 40 ms), i.e. decreasing the bandwidth of each sub-band. The non-overlapping period of the frames is called the hopping time. The overlapping between each two-adjacent overlapping-windowed frames is 8 to 40 ms. The hopping time of that overlapping-windowed frame is 8 to 20 ms. The research of this thesis uses the following well-knowing windows: Hamming, Hanning and Blackman [1, 2].

The first step of standard audio and speech-DSP sequence is converting the analogue signal to discrete samples. Encoding of the discrete samples to their approximated equivalent values (the quantization) is the second step. To reduce errors of that approximation, suitable quantization levels are chosen. According to the nature of the human ear, 8 bits (256 levels) is enough to hear audio and/or speech as a deterministic audio and/or speech signals. 16 bits ($\approx 32 \times 10^3$ quantization levels) make those signals highly-defined audio and speech. For the audio signals (e.g. music), their definitions are very important, so number of bits either 16, 24 or 32 (for the highest definition, levels/sample $\approx 4 \times 10^9$ for 32 bits) [1, 2].

Statistical parameters of the speech signal are slightly changed for the same speaker when he/she says same words/sentences and/or from time to other times. Due to that parametric changing, the accepted stochastic description of the speech is quasi-stationary signal. Such stochastic property of the speech signal, because there are no guarantees to estimate the chances (the certain probabilities) of the events of the speech signal perfectly compared with the non-event occurrences [1, 2, 4]. The statistical distribution of speech signal is a dynamic stochastic process in time, frequency and other parametric domain, due to the unpredictable changes in the speech Model-State. The state-dependent probability distribution (s.d.p.d.) of a speech sequence can take a variety of forms [4]. When the speech frame has known number of samples N_w and they are continuously distributed

components, usually its statistical distribution is the same N_w number of dimensional Gaussian distribution (or a mixture model of such distributions). Although they are not widely used, some other models are considered in the literatures, supposed that the s.d.p.d. depends on the current state, the previous state and the previous observation [1, 2, 4, 5].

1.3 Pitch of Speech Signal

According to Fourier Transform (FT), frequency domain representation of any periodical waveform is discrete complex values. Magnitudes of these complex values is the frequency components of that waveform. The first component is called fundamental-frequency. The 2nd, the 3rd component, ... etc. are called the 2nd, the 3rd harmonic ... etc. [1, 2].

There is somehow similarity between waveforms of most speech signals and the typical Amplitude Modulation (AM) waveforms. Approximately, carrier tones of the speech waveform locate in the range of 40-4000 Hz. Frequency deviation of the speech virtual carrier is several tens Hz. Such waveform has somehow of periodicity (semi-periodicity) [1, 2]. According to the above Fourier Transform analysis, frequency domain spectrum of the speech signal has somehow of discrete-value sub-bands (see Figure 1.1 [6]. Instead of the discrete unit-impulses of each harmonic for the full periodical signal, speech signal has discrete groups. Each group consist of adjacent unit-impulses. Approximately, gradient of weights of those unit-impulses are bell-shaped Gaussian distribution. Rounding estimation (by regression and/or interpolation techniques) of amplitude distribution of each group is a normal Gaussian distribution (Figure 1.1). Since the first harmonic is called the fundamental frequency of the periodical signal, the first group of speech signal is called the Pitch [6-8].

Gaussian Mixture Modelling (GMM) could formulate the mathematical model of such groups. Each group has its unique average value (centroid) and unique variance value. inside each group, statistical probabilities for transitions of each sub-harmonic to their adjacent sub-harmonics could be mathematically modeled by Hidden Markov Modeling (HMM). Also, statistical probabilities for transitions from each group to its adjacent groups could be mathematically modeled by HMM and GMM using its sufficient statistical database.

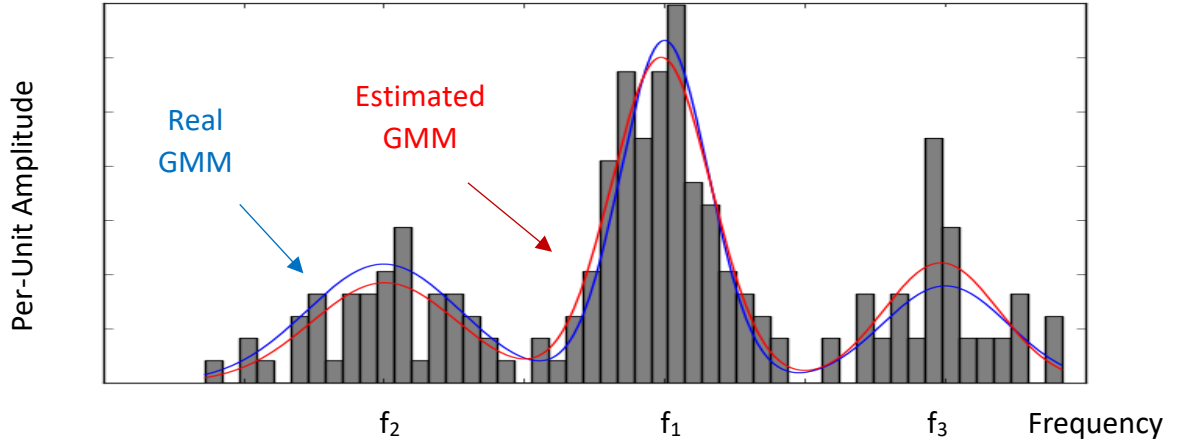


Figure 1.1 Typical sketch of statistical distribution of the harmonics for long time of speech. The time should be 20 to 30 minutes. Each group is an estimated & a real Gaussian distributions. All of them are GMM modeled.

Pitch concept is a key to solve a lot of speech-DSP problems (e.g. speech separation). The first well-known algorithm to estimate the pitch location is by *Noll* in 1967 [7]. During the past decades, many algorithms have been experimented to estimate centroids of the pitches under the title Pitch Detection Algorithm PDA. For this research, Robust Algorithm for Pitch Tracking (RAPT) [8] is used many times to estimate the pitches and to remove the unvoiced speech. Frame-by-frame, Pitch locations are founded for each effective 8 to 16 ms of speech frame. The locations are in the range 40 to 600 Hz [7, 8].

Experimentally, spectrum of the speech is a result of 16 to 40 ms Short-Time Fourier-Transform (STFT). Effectively, 8 to 16 ms of hopping time, the spectrum is hopping from each analyzed frame to its next adjacent frame. To represent specific speech segment entirely, both time and frequency domains should be considered in one graph. For that, the all-frames STFT transformations of speech are arranged where the time domain in horizontal-axe of the graph and the frequency domain in vertical-axe. The graph is called the Spectrogram (see Figure 1.2). Behavior of speech is linear in time domain but non-linear (logarithmic) in frequency domain. For those, horizontal-axe is plotted using the linear scale and vertical-axe is plotted using the log-scale. Since the magnitude and/or the amplitude strength of each sub-band have wide range, the linear scale of them is non-sensible. For the sensible measurement of that strength, decibel (dB) unit is calculated [7, 8].

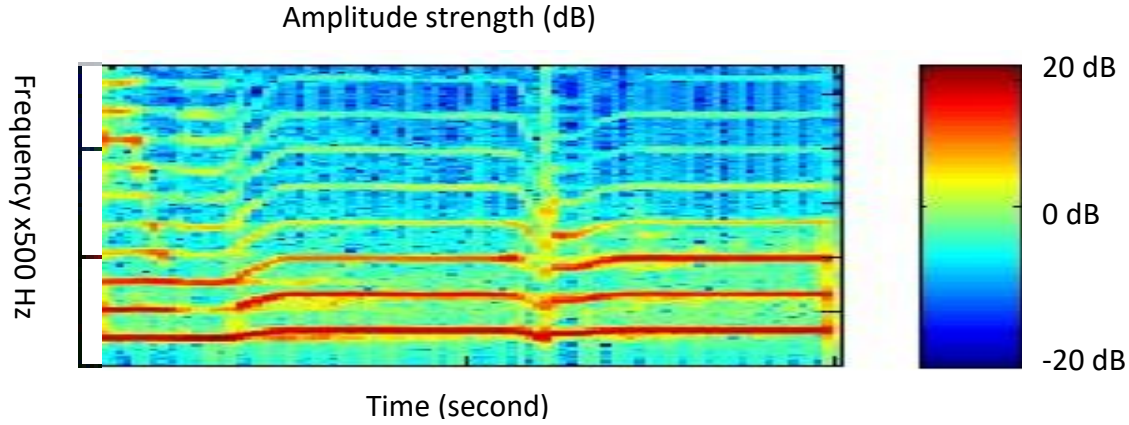





Figure 1.2 Typical STFT Time-Frequency Spectrogram for 2 second of speech. The 1st lower red line is the pitch of sequential frames. The others red and yellow lines are the harmonics.

1.4 Spontaneous Conversation, Dialog Speech and Mixture Speech

A conversation is the oral chat among multi-speakers. Spontaneous conversation contains two formats of speech, dialogue and mixture (is called overlapped-speech). In this thesis, the speakers are two: Female (F) with Female (F) (its symbol is FF), Male (M) with Male (M) (its symbol is MM) and Female (F) with Male (M) (its symbol is FM). To clarify that, let Female F with Male M are speaking in a spontaneous conversation (see the (a)/Figure 1.3). Graphically, speech segment of F alone is sketched as vertically-lined rectangle  and speech segment of M alone is sketched as horizontally-lined rectangle . Speech segment of F with M simultaneously (FM) is sketched as mesh-lined rectangle . Dialog speech is a format of speech between the two speakers when only one of them e.g. F is talking and other (M) is silent. During dialogue speech periods, another speaker M is not talking (when F is talking, M is silent and vice-versa). Mixture speech FM is a format of speech between the two speakers when the speakers (F and M) are talking simultaneously.

A spontaneous conversation contains dialogue and/or mixture speech formats. Sometimes, one speaker is talking and other times the two speakers are talking. The first waveform of Figure 1.3 shows an illustrative arbitrary spontaneous conversation of F with M speakers. The conversation is supposed as a sequence of nine speech segments. First segment is a mixture of F with M, then the second segment is a dialogue where F continues alone, then the third segment is a dialogue

where M is talking but F is not, then the fourth segment is a mixture where M continues with F simultaneously, then the fifth segment is a dialogue where M continues alone, then the sixth segment is a mixture where M continues with F simultaneously, then the seventh segment is a dialogue where F continues alone, then the eighth segment is a dialogue where M is talking but F is not, and the last (the ninth) segment is a mixture where M continues with F simultaneously.

During that spontaneous conversation, switching instant from any speech format to another speech format is the transferring moment from dialogue speech to mixture speech or vice-versa. In fact, duration of a dialogue or a mixture speech is random process, because it depends about the speakers' personality, the conversation situation, subjects of the talking and the circumstances of the conversation session. For example, during a regular presentation, period of dialogue format may be more than several times period of mixture format. In contrast, an interruption of overlapped-speech (mixture format) by one word (e.g. yes) consumes fraction of second. Sometimes, that short-period of speech is insensible or inaudible, so it could be negligible [9].

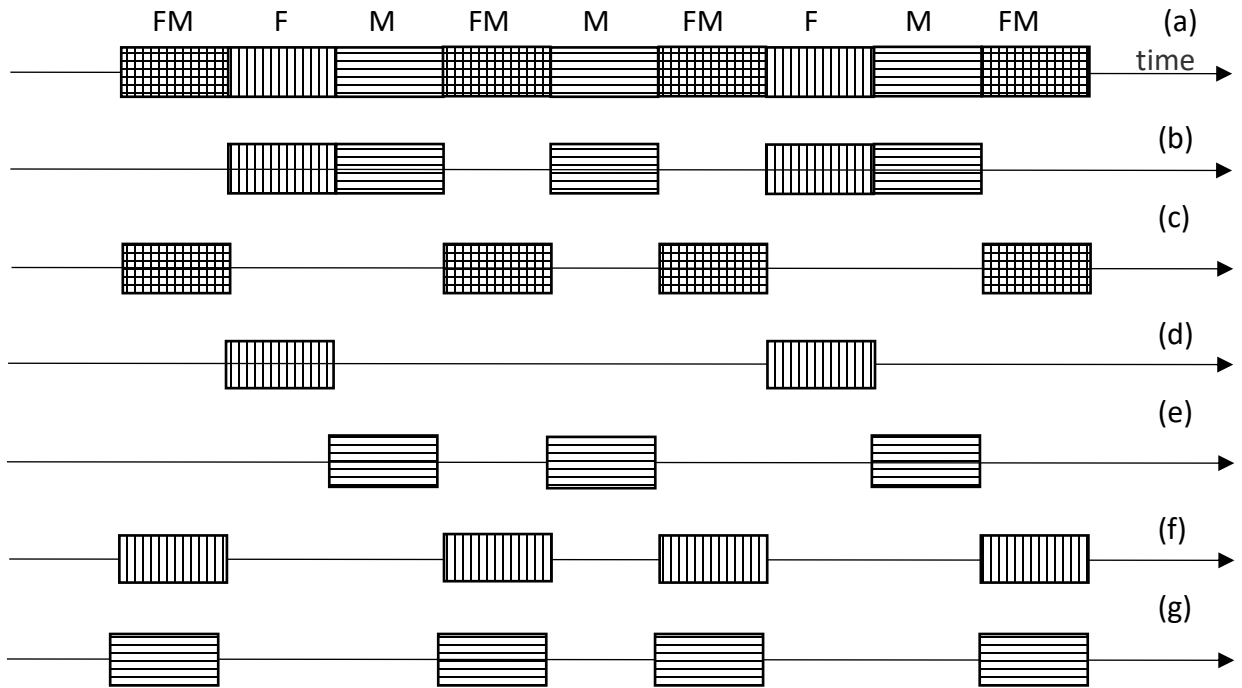


Figure 1.3 Arbitrary spontaneous conversation. The dialogue Female F (vertically-lined) alone with Male M (horizontally-lined) alone. The mixture FM is both simultaneously (cross-lined). Input and outputs signals of: overlapped-speech detection (the (a), the (b) and the (c)), speaker diarization (the (b), the (d) and the (e)) and speech separation (the (c), the (f) and the (g)). N.B. There are horizontal-axes time-domain relationships between all the sketches.

1.5 Overlapped-Speech Detection, Speaker Diarization and Speech Separation

Finding those switching instants from any speech format to the another is the solution key to segregate the two formats of spontaneous conversation. Overlapped-speech detection tries to estimate those switching instants. Efficient estimation method segregates the mixture speech form the dialogue speech properly [9, 10]. Input signal of the overlapped-speech detection is a spontaneous conversation (e.g. the (a)/Figure 1.3). The outputs of the overlapped-speech detection, are two segregated speech signals: dialogue (the (b)/Figure 1.3) and mixture (the (c)/Figure 1.3).

For the dialogue speech, speech-DSP utilizes speaker diarization method to isolate the individual speech segments of each speaker. Speaker diarization consist of two main phases: speaker segmentation and speaker clustering. Hierarchical Clustering Scenarios (Top-Down Divisive and Bottom-Up Agglomerative Scenarios) support clustering phase [11, 12]. Input signal of the speaker diarization is a dialogue speech signal (e.g. the (b)/Figure 1.3). The outputs of the speaker diarization are two isolated speech signals: Female F (the (d)/Figure 1.3) and Male M (the (e)/Figure 1.3) [13, 14].

For the mixture speech, speech-DSP utilizes speech (source) separation algorithms to separate the mixture into its original speech of each speaker (they are called targeted speech) [15, 16]. Input signal of the speech separation is a mixture speech signal (e.g. the (c)/Figure 1.3). The outputs of the speech separation are two separated speech signals: Female F (the (f)/Figure 1.3) and Male M (the (g)/Figure 1.3).

1.6 Samples, Window-Frame and Hopping period

The input observation signals of this thesis chapters are: spontaneous conversation speech, dialogue speech format and mixture speech format. These formats are collection of speech segments. The conversation consists of sequential of these segments (Figure 1.3 is an example for arbitrary sequence of such conversation with their segments of speech formats). The following definition for the processed durations are necessary in the thesis description [1, 2].

Let the total length of the conversation is T_t seconds. The conversation is converted from its original analogue form to the equivalent discrete samples. Total number of these samples is N_t for the all conversation. Sampling rate of the discrete form is f_s sample/s:

$$\text{Nunmer of samples} = \text{Duration} \times \text{sampling rate} \quad (1-1')$$

$$N_t = T_t \times f_s \quad (1-2)$$

where f_s is 8000 or 16000 samples/s in the research. The speech is processed by dividing its signal into number of overlapping-windowed frames. Each frame is scaled by standard window e.g. Hanning window.

The scaling focuses on the center of the window, with the gradually attenuation for the bilateral borders of the window. Duration of the window T_w is in the range 16 to 40 ms. For the Fast Fourier Transformation (FFT) of the speech window from time domain to frequency domain, T_w should be radix-2, i.e. 8, 16, 32 or 64 ms. Number of the samples for each overlapping-windowed frame N_w is:

$$N_w = T_w \times f_s \quad (1-3)$$

According to (1-3), number of samples N_w is 64, 128, 256, 512 or 1024 samples. FFT transforms those overlapping-windowed frames to frequency domain, so FFT point is N_w per frame. Each transformed frame has N_w complex-value points in frequency domain. Each i^{th} element of the frequency domain frame is complex-conjugate of the $(N_w - i^{\text{th}})$ element. For the frame, this property is called the mirror-conjugate. The i index is from 1 to $N_w - 1$. This property does not include the first and center elements, where they are real values. For that, the considered number of the complex values of the frequency domain is $(N_w/2)+1$. This number is number of the Filter-Bank sub-bands. Absolute value (magnitude) of them is the sub-band content.

The standard overlapping ratio of any frame with its adjacent Left-Hand-Side (LHS) or Right-Hand-Side (RHS) is 67% to 50% of the window width. The non-overlapping period is called the hopping period T_h . Ratio of that period is 33% to 50% of the window width. Number of the samples for each hopping period N_h is:

$$N_h = T_h \times f_s \quad (1-4)$$

According to (1-4), number of samples N_h is 40 to 256 samples. That duration is important for speech DSP processing. It should be in the range 8 to 20 ms. Since the conversation consists of N_t samples and each hopping period is N_h , total number of the processed frame N_f is:

$$N_f \approx \text{floor}\left(\frac{T_t}{T_h}\right) \approx \text{floor}\left(\frac{N_t}{N_h}\right) \quad (1-5)$$

where, $\text{floor}(\cdot)$ is the down-rounding floor function. The resulting spectrogram of the Time-Frequency TF domain is $[((N_w/2)+1)\text{-by-}N_f]$ matrix [1, 2].

1.7 Supervised, Semi-Supervised and Unsupervised Machine Learning

Attributes of DSP and Machine Learning (ML) system are: observation signal(s), database and/or useful information. When ML system has only observation signal(s), it is called Unsupervised ML (i.e. does not have any database and/or useful information). When the Unsupervised ML system does not have database and/or useful information, but it can generate any useful database and/or useful information during the process, the system is called Semi-Supervised ML. In addition to the observation signal(s), Supervised ML system has prior database and/or useful information, to support the required process. The Figure 1.4 illustrates those main definitions and categorizing of ML systems [17-19]. Observation signal(s) are the input signal(s) which are processed by the main DSP approach. Database and/or useful information are useful information which have relationships with the observation signal(s). Observation signal(s) are audio, speech, video, image, object... etc. signals or composite signal(s) from them.

For a specific observation signal(s), database and/or useful information should be on the same area of those observation signal(s) or have a useful relationship with them (e.g. speech signals, audio features, sub-bands energy and/or other related parameters or characteristics).

For this research, input observation signal is a spontaneous speech of two speakers. The TIMIT audio and speech library has been used to generate some of the required observation signals and database [20]. The library is a reliable source for audio, speech and speakers. Carefully, (*FileName.wav*) speech files are chosen from that library. The library avails the required speech narration for 20 to 30 minutes of each one of two speakers (Female F and Male M). The speech is digitally-recorded using the highest audio definition, i.e. 44100 sample/s. Each sample has the highest number, 32 bit/sample. For this research, in addition to the recorded speech of TIMIT library, speech of 30 speakers are prepared [20]. The speakers are female and male narrators of audio books. The definition of their speech files is deterministic using 8000 and 16000 sample/s sampling rate with 16 bit/sample resolution. The (*FileName.wav*) speech file of each speaker contains about one hour of continuous narration speech.

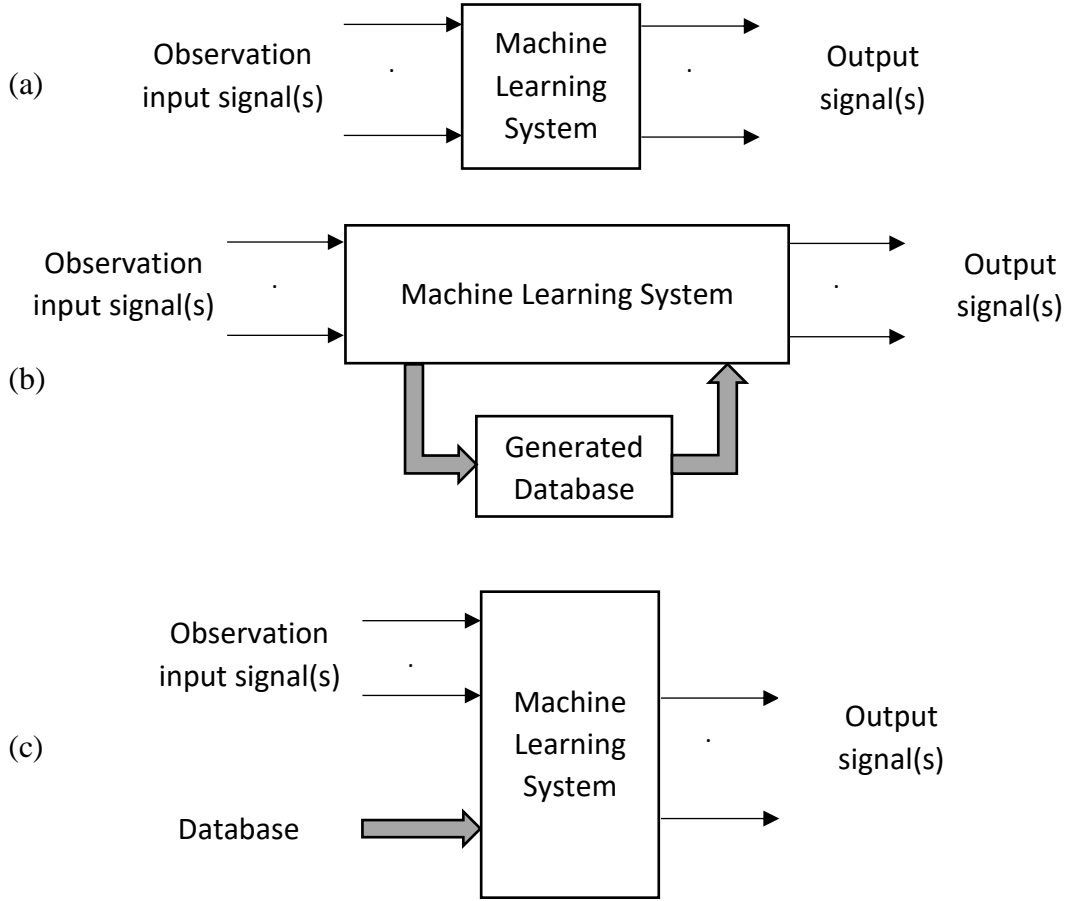


Figure 1.4 Typical machine learning DSP systems. The 1st is Unsupervised, the (b) is Semi-Supervised and the (c) is the Supervised machine learning.

1.8 Blind Speech Separation versus Informed Speech Separation

Input observation signal(s) of speech separation is a mixture speech of multi-speakers. For this thesis, the research is focusing on single channel speech separation where input observation signal is only one signal. For this thesis, number of speakers per each conversation is two. Generally, the speakers of mixture speech are speaking simultaneously for a specific period. When the attribute of the separation is only the mixture speech, the process is unsupervised machine learning which is called Blind Speech Separation. The traditional mathematical-based approaches for the speech separation are: Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Non-negative Matrix Factorization (NMF). Computational Auditory Scene Analysis (CASA) simulates the human ear behaviour to follow and separate speech of multi-speaker conversation. Almost, speech separation process attempts to separate components (e.g. the pitches) of the Time-Frequency pattern of the mixture speech.

In contrast, the speech separation is semi-supervised or supervised Machine Learning when a database is used in the processing. It is called Informed Speech Separation [15, 16, 21]. The approaches for that are: video-assisted separation, the spatial object coding, reverberant models for source separation, score-informed source separation, language-informed speech separation, user-guided source separation, source-separation informed by cover version and informed source separation applied to speech, music or environmental signals.

1.9 Overall-System

Overall-system input observation signal is a spontaneous conversation speech, i.e. contains dialogue speech and mixture speech (called overlapped-speech) of two speakers. The main aim of the research is the isolation of the individual speech signal of each speaker from that conversation. The Figure 1.5 and Figure 1.6 show two proposed schemes to achieve those tasks. The first proposed scheme is an unsupervised Machine Learning system. The second scheme is a semi-supervised Machine Learning system.

The first system (see Figure 1.5) consists of overlapped-speech detection block at first, to segregate the dialogue and the mixture speech signals (more details in Chapter 3). Speaker diarization block isolates the speech segments of each speaker. Blind speech separation block separates the speech segments, but could not identify the speaker of any segment. Speaker diarization could identify the speaker of each segment by speaker clustering method. For the blind speech separation, more details in Chapter 4 [22].

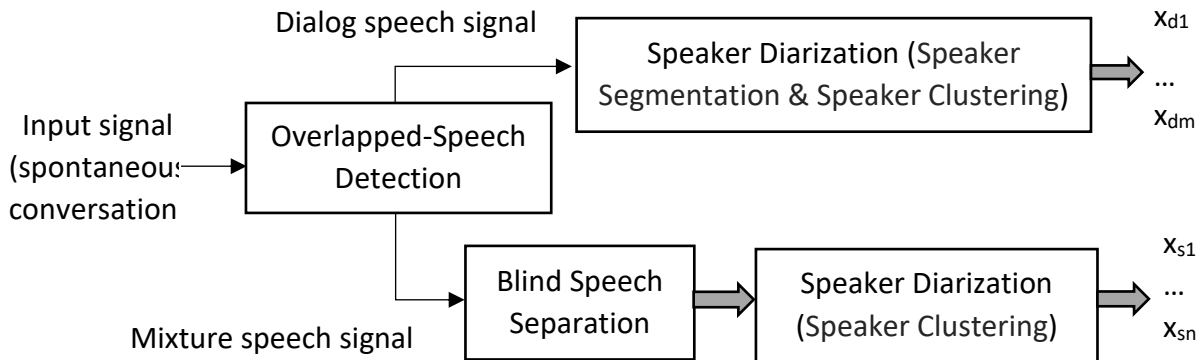


Figure 1.5 Chapter 4 overall system. The input is spontaneous conversation signal and the outputs are the individual speech signal of all the speakers. The system is an unsupervised Machine Learning system.

The second system (see Figure 1.6) consists of overlapped-speech detection at first, to segregate the dialogue and the mixture speech signal (more details in Chapter 3). Speaker diarization block isolates the speech segments of each speaker. The Informed speech separation block separates the speech segments and identify the speaker of the segments [23].

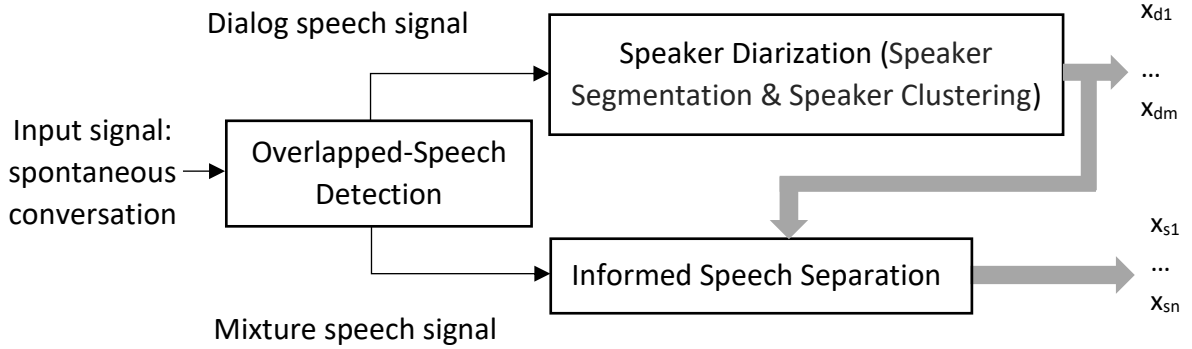


Figure 1.6 Chapter 5 overall system. The input is spontaneous conversation signal and the outputs are the individual speech signal of all the speakers. The system is semi-supervised Machine Learning system.

The system is semi-supervised machine learning because the speech separation process is informed, where the system exploits the output information of the speaker diarization block. Speaker diarization could identify the speaker of each segment. For the Informed speech separation, more details in Chapter 5. For the detection, the diarization, the separation and the overall-system, typical sketches of their inputs and outputs waveforms are illustrated in Figure 1.7, Figure 1.8, Figure 1.9 and Figure 1.10 respectively.

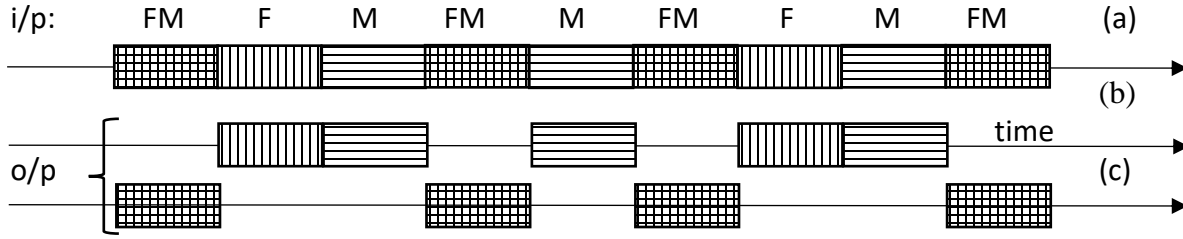


Figure 1.7 Typical sketches of the overlapped-speech detection. The (a) is the input (i/p) signal of the system. It's the i/p of the overlapped-speech detection. The (b) is the 1st output (o/p), the dialogue speech, the i/p of speaker diarization. The (c) is the 2nd o/p, the mixture speech, the i/p of speech separation. There are horizontal -axes time-domain relationships between the sketches.

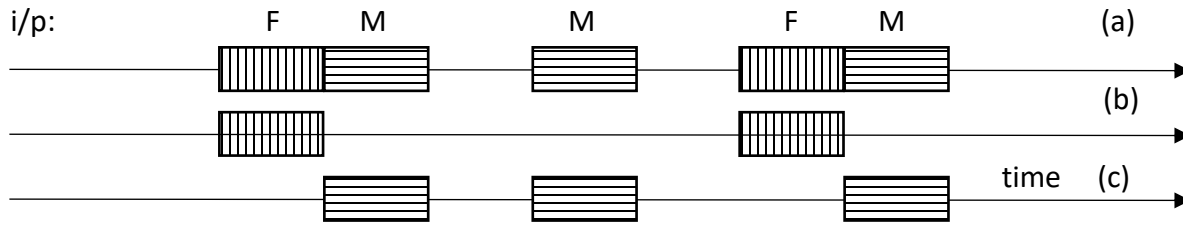


Figure 1.8 Typical sketches of the speaker diarization process. The (a) is the i/p of the speaker diarization block (dialogue speech). The (b) is the 1st o/p, the Female F speech. The (c) is the 2nd o/p, the Male M speech. The output is the input of the Informed Speech Separation.

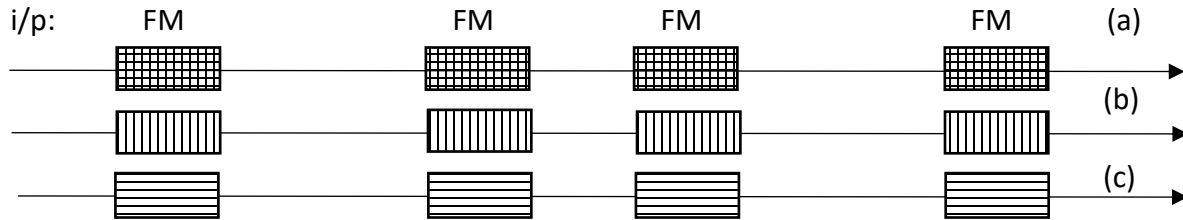


Figure 1.9 Typical sketches of the speech separation process. The (a) is the 2nd o/p of the speaker segregation, the i/p of the speech separation (mixture speech). The (b) is the 1st o/p of the speech separation, the Female F speech segments. The (c) is the 2nd o/p of the speech separation, the Male M speech segments. There are horizontal -axes time relationships between the sketches.

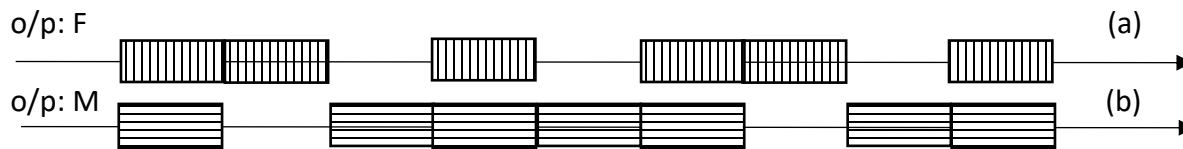


Figure 1.10 Typical speech-DSP output signals for the spontaneous conversation. The (a) is the female F output speech signal, from the overall system, collected from the F of speaker diarization and speech separation blocks. The (b) is the male M output speech signal, from the overall system, collected from the M of speaker diarization and speech separation blocks.

1.10 Subjective Test versus Objective Test

To check the efficiency of a DSP algorithm, its actual output signals are compared with the theoretical derived or measured output signals. In machine learning and pattern recognition, there are well known measuring distances to check the efficiency level of these algorithms (e.g. Euclidean distance).

There are two methods of checking the output speech, subjective and objective tests. During the experiments simulation, there are a lot of implementation trials. After each trial, their output speech could be tested by suitable listeners. The listening conditions should be attended carefully. The listeners hear the resulting speech required times in quite conditions. Each listener has his/her individual ability to assess the output speech, and compare it with the original reference speech. This method of test is called subjective test [24-26]. Subjective test needs enough number of listener to satisfy the statistical requirements. The resulting assessment is the average value of all assessments. At last, variance of these assessments is taken in the researcher consideration. The smallest variance is the best assessment and the highest is the worst. That step is essential to elaborate the crude assessments, then produce the net assessment [24-26]. In fact, evaluation of the output speech using subject tests is not reliable. The alternative reliable checking is the mathematical-based objective test [24-26].

The standard objective test of the speaker diarization and the overlapped-speech detection is Diarization Error Rate (DER). There are three overlapped-speech detection errors:

- Missed-Speech Error (E_{MISS}): When the detection suggests a mixture speech as a dialogue speech.
- False Alarm Rate (E_{FA}): When the detection suggests a dialogue speech as a mixture speech.
- Overlap Speaker Error (E_{OVL}): The overall error of the overlapped-speech detection.

The tests are Per-Unit (PU) or Percentage % ratio of error speech to total speech. For free-of-error output speech, Per-Unit test rates are 0.0 (Percentage is 0%). For full-of-error, Per-Unit test rates are 1.0 (Percentage is 100%) [27]. The test can evaluate the error of one speaker segments (e.g. F segments) compared with the reference original speech of that speaker. Almost, the tests can rate the errors of the entire conversation segments for all the speakers. Per-Unit DER is:

$$DER = \frac{\sum_{s=1}^S dur(s) \cdot \left(\max \left(N_{ref}(s), N_{hyp}(s) \right) - N_{correct}(s) \right)}{\sum_{s=1}^S dur(s) \cdot N_{ref}} \quad (1-6)$$

where, the total number of speaker segments is S . Both reference (original) and hypothesis (output) signals contain the same speakers; it is obtained by collapsing together the hypothesis and reference speaker turns. $N_{ref}(s)$ and $N_{sys}(s)$ indicate number of speaker speaking in segments. $N_{correct}$ is the number of speakers those speaking in s [27]. The DER tests have been used in Chapter 3 to evaluate the output speech segments of the first speaker F, the second speaker M of each conversation and for all that conversation.

For the speech separation, the reference speech is called the targeted speech. There are several speech separation objective tests [25, 28]. The most reliable tests are the following four tests:

- The energy Source to Distortion Ratio SDR test (by the decibels dB) is:

$$SDR (dB) = 10 \log_{10} \frac{\|S_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \quad (1-7)$$

where, S_{target} is the reference signal, e_{inter} is the interference error, e_{noise} is the noise error and e_{artif} is the artifact error. The error is the absolute distance between the output signal and the reference targeted signal.

- The energy Source to Interferences Ratio SIR test is:

$$SIR (dB) = 10 \log_{10} \frac{\|S_{target}\|^2}{\|e_{interf}\|^2} \quad (1-8)$$

- The energy Source to Noise Ratio SNR test (by the decibels) is:

$$SNR (dB) = 10 \log_{10} \frac{\|e_{interf} + S_{target}\|^2}{\|e_{noise}\|^2} \quad (1-9)$$

- The energy Source to Artifacts Ratio SAR test is:

$$SAR (dB) = 10 \log_{10} \frac{\|e_{interf} + e_{noise} + S_{target}\|^2}{\|e_{artif}\|^2} \quad (1-10)$$

To evaluate algorithms of Chapter 4 and Chapter 5, the SAR, SDR and SIR tests have been used.

1.11 Masking

The last step in the overlapped-speech detection, the speaker diarization and the speech separation jobs is the identification of the processed parameters of the speech and the speakers. The identification belongs the desired parameters to specific speech and speakers. Also, that identification avoids the undesired parameters from those speech and speakers. In machine learning and pattern recognition DSP, such task is listed under the statistical classification and clustering title [18].

The classification and clustering methods discriminate between the desired and the undesired signals (e.g. speech signal). Also, those methods discriminate between the desired and the undesired components (e.g. audio features). In overlapped-speech detection, speaker diarization and speech separation, the classification and clustering tasks are achieved by masking techniques. The main job of the mask is the final identification decisions. The mask magnifies the desired signals and components, and attenuates the unwanted signals and components.

There are two masking strengths: the binary masking and the soft masking. To choose the best masking strength, experimental results of the classification and clustering methods are used. When the classification and clustering indicate that the discrimination and decision are sharp, the binary mask nulls the non-desired labels, and then shares all the parameters to the desired label entirely. When the classification and the clustering indicate that the discrimination and decision is not sharp, the soft mask shares the parameters to the desired labels proportional with the distances (e.g. Euclidian distances). The mask makes the decisions in time domain, frequency domain and on the audio features as well [29, 30].

For the speaker diarization process, the binary mask is used. The observation signal of the speaker diarization is a dialogue conversation between multi-speakers, where only one speaker is speaking, while other speakers are silent. At first, the speaker diarization divided the overall observation speech signal to specific number of speech segments. Each segment belongs to one unknown speaker only. In the speaker diarization, this phase is called speaker segmentation. In the speaker

diarization, to identify the speaker of each segment, the speaker clustering is used. Speaker clustering is the second phase of the speaker diarization. According to the distances between the audio features of the speakers, the clustering identifies each segment. Mapping of the identification is one-to-one transformation. Such mapping should be implemented by the binary mask, i.e. entirely, any one of those segments must be classified for one-and-only-one speaker [18, 29, 30]. In this thesis, there are two speakers (F and M) involve with a dialogue format speech. Parametric model represents the time and the frequency domains and the audio features of that speech. Let the distance from F parameters to a reference-parameters is d_f , and the distance from M parameters to a reference-parameters is d_m . The share is calculated by taking the relative inversely relation of those distances. According to the binary mask the shares of F and M are:

$$\left. \begin{aligned} F \text{ share} &= \text{all the parameters, if } d_m > d_f \\ M \text{ share} &= \text{all the parameters, if } d_f > d_m \end{aligned} \right\} \quad (1-11)$$

The above masking technique has been adopted for the Chapter 3 algorithm. Overlapped-speech detection algorithms set is a subset of the diarization algorithm set. Chapter 3 algorithm meets the conditions of the binary masking (more details in Chapter 3).

Some cases in the speech separation, the binary mask have good ability to complete the job. Other cases in the separated speech, an output of one speaker contains parameters belong to the other speakers. For this reason, almost the sharp decisions of the binary mask add speech parameters of a speaker to other speakers. In this situation, the other speakers have lost part of their audio parameters. For any speaker, the accumulated errors are the additive parameters of the other speakers, and the lost parameters of that speaker. In this case, the soft mask is the alternative because it has the right decisions to divide the parameters among the speakers.

For the two speakers: F and M, there are two numerical calculation methods for the soft masking:

1. The first method is by taking the amplitude ratios of the distances, i.e.:

$$\left. \begin{aligned} F \text{ share} &= (d_m / (d_f + d_m)) \text{ of the parameters} \\ M \text{ share} &= (d_f / (d_f + d_m)) \text{ of the parameters} \end{aligned} \right\} \quad (1-12)$$

2. The second method is by taking the energy ratios of the distances, i.e.:

$$\left. \begin{aligned} F \text{ share} &= ((d_m)^2 / ((d_f)^2 + (d_m)^2)) \text{ of the parameters} \\ M \text{ share} &= ((d_f)^2 / ((d_f)^2 + (d_m)^2)) \text{ of the parameters} \end{aligned} \right\} \quad (1-13)$$

For the blind speech separation algorithm in Chapter 4, the binary and the soft masks are used. For the soft mask, the two types of masks (the amplitude calculation and the energy calculation) are used. The performances of these three masks are fluctuating (more details in Chapter 4).

For the informed speech separation algorithm in Chapter 5, the soft and the binary masks are used. Optimization arrangement improves the fluctuating effect of the soft and binary masks (more details in Chapter 5).

1.12 Objectives and Aims of the Research

The observation signal of the research is a speech signal of spontaneous conversation of two speakers (e.g. the (a)/Figure 1.11). The objectives of the research are the extracting of the individual speech of each speaker alone. The (f) and the (g)/Figure 1.11 are the objectives for the (a)/Figure 1.11 conversation. They are the individual isolated speech of the two speakers, each one alone.

The main aims of the research are:

1. The overlapped-speech detection of the input signal. The (b) and the (c)/Figure 1.11 for the (a)/Figure 1.11 conversation. Novel algorithm has been contributed in the research.
2. The speaker diarization of the dialogue output of the overlapped-speech detection. The (b) and the (c)/Figure 1.11 for the (a)/Figure 1.11 conversation. Existing toolbox is invoked in the research.
3. The speech separation of the mixture output of the overlapped-speech detection. The (b) and the (c)/Figure 1.11 for the (a) /Figure 1.11 conversation:
 - 3.1 Without utilizing the outputs of the speaker diarization (blind speech separation). Novel algorithm has been contributed in the research.
 - 3.2 With utilizing the outputs of the speaker diarization (informed speech separation). Novel algorithm has been contributed in the research.
4. The following secondary aims are exploited: the k-means clustering, the filter-bank analysis and synthesis, the non-negative matrix factorization, the RASTA-PLP audio features extraction, the hierarchical clustering scenarios (top-down divisive and bottom-up agglomerative) and the speaker clustering.
5. The basics and the principles of: the machine learning, the pattern recognition, the random variables (the statistics) and the random process (the stochastic), have been based-on in a wide-range, deeply and continuously. To support these terms, measuring by pattern

distances are used. The well-known basics and the principles of speech-DSP are used to get the efficient and to avoid the inefficient: algorithms, algorithms and parameters. The subjective and objective tests have been used to check and evaluate the resulting speech.

The reliable software has been invoked to simulate the existing and the novel: algorithms, techniques, toolboxes and algorithms, to investigate the truthfulness of them.

To adopt the suitable audio features, RASTA-PLP, MFCC and PNCC have been compared. Choosing the k-means is decided after several comparisons between several clustering algorithms. For the speaker diarization of the dialogue speech, several toolboxes are tested. Several teens of narrators are downloaded to pick up the best 33 narrators. To choose the better parameters, most possibilities are take into consider.

1.13 Contributions

The research proposes three main contributions:

1. The algorithm “overlapped-speech detection based-on stochastic prosperities”. The algorithm contains the novel concept “Grouping” to optimise collection of audio features instead of optimise the features themselves. The algorithm based-on the stochastic prosperities of the PDF of the clustered audio features. The algorithm exploits basics and principles of machine learning and pattern recognising. k-means clustering has been used several times to draw required borders of the thresholds. The interaction between the hierarchal scenarios, the k-means clustering and the optimised groups has been used to finalize the algorithm efficiently.
2. The algorithm “blind speech separation by filter-bank, non-Negative matrix factorization and speaker clustering”. The algorithm uses speech frames instead of the segments to avoid the possibility of speaker segmentation error. To reduce the resulting errors of the traditional speech separation, at first the algorithm separates the speech but does not identify the speaker of that speech. Speaker clustering complete that job by identify the speaker efficiently. The sub-signals of the filter-bank are factorised by the NMF instead of the tradition NMF of the main signal.
3. The algorithm “informed speech separation by semi-supervised non-Negative matrix factorization”. The outputs of two parts of the research (the detection and the diarization) are used to target the original speech components of a spontaneous conversation. The

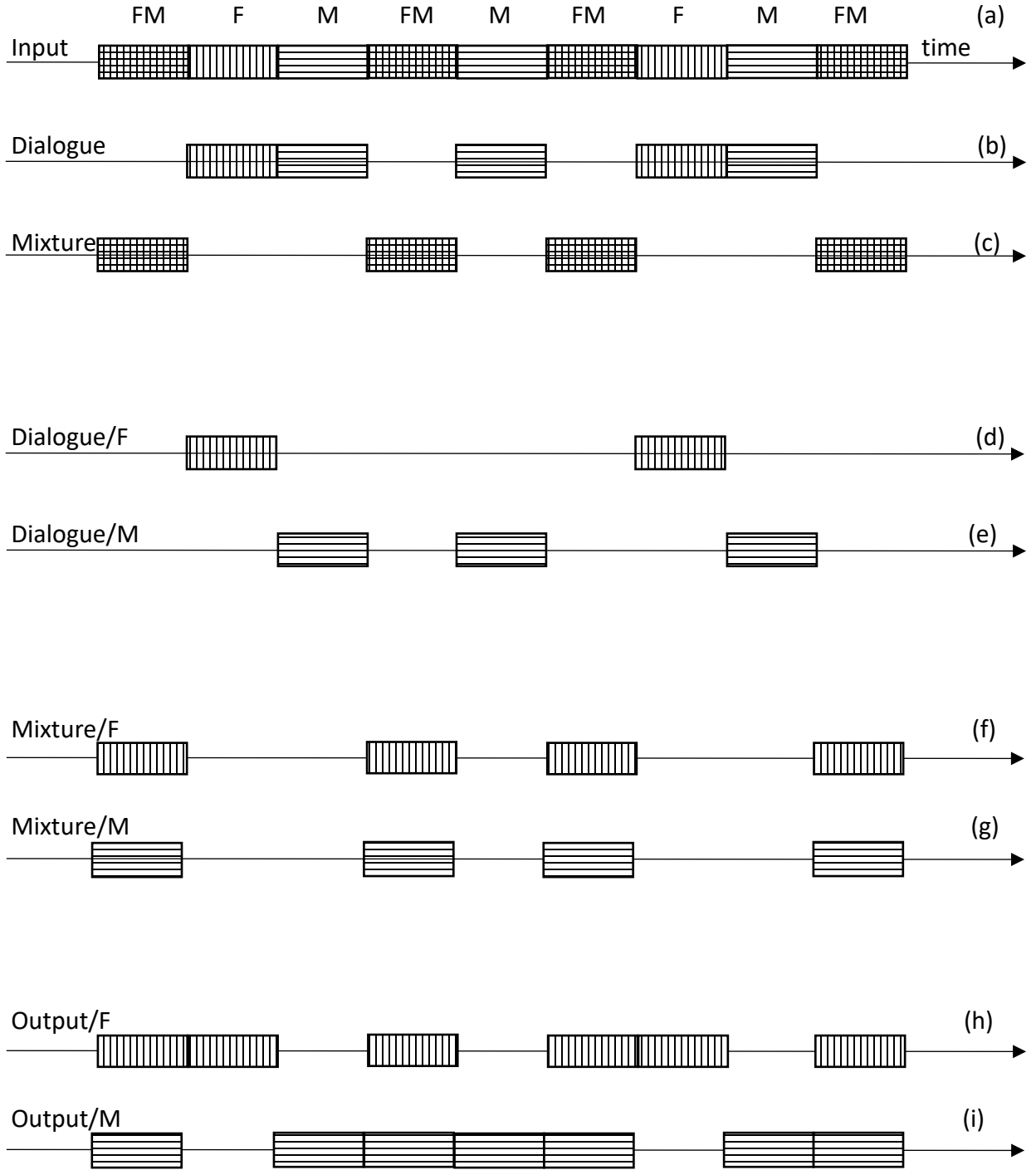


Figure 1.11 Typical sketches of Aims and Objectives of the research. The (a) is the input observation signal (spontaneous conversation between two speakers: F & M). The (b), the (c), the (d) and the (e) are the Aims of the research. The (f) and the (g) are the Objectives of the research. The (h) is the individual speech of F. The (i) is the individual speech of M. There are horizontal-axes time-domain relationships between all the sketches.

training of NMF of the generated data-base is configured virtual mixture speech instead of the isolated data-base of the speech. For this algorithm, last contribution is the simultaneously soft and binary masking. The algorithm is dynamically switching from mask to the another, according to the best performance.

The research covers the complete system that represents a spontaneous conversation. The research tackles the different problems of that conversation. In fact, there is no conversation contains mixture or dialogue speech format only. In fact, the actual conversations contain both.

Chapter 2. Literature Reviews

2.1 Introduction

The chapter survives several recent years' background of the main three achievements of the thesis (overlapped-speech detection, blind speech separation and informed speech separation) including: the papers, the thesis, the software, the reports, the presentations, the patients and the corpuses.

At first the survey prefaces the scope of the overlapped-speech detection algorithm which will be the material of Chapter 3. The survey scans the most important literatures during the past 10 years. Since the detection of the overlapped-speech is subset of the speaker diarization, the survey involves also the literatures of the diarization which related to the detection. The recent available speaker diarization corpuses are used to execute the current missions of the overlapped-speech detection, because there is no dedicated corpus for the detection only. The survey covers the most important corpuses during the past 10 years. The rich several theses about the diarization are regarded also.

Then, the survey prefaces the job of the blind speech separation algorithm which will be the material of Chapter 4. The survey scans the most important literatures during the past several years also. The algorithm is done by the cooperating of the filter-bank technique, the NMF technique and the speaker clustering. The survey focuses on the common literatures for these approaches. The survey is very important for the final cross-checking between the algorithm approach and objective tests, with the approaches and tests of those literatures. Because the achieved NMF-based speech separation is seldom, the NMF-based audio and sound separation is involved.

The survey prefaces the job of the Informed speech separation algorithm which will be the material of Chapter 5. The survey scans the most important literatures during the nearest years. The algorithm is done by the using of the semi-supervised machine learning principles. Almost the achievements of the NMF-based separation are applied on the audio and the sound rather than the speech. This is the main problem for the speech separation which is based on the NMF. The semi-supervised NMF-based separation is subset of other informed speech separation frameworks. To increase the chances of the cross-checking, those other frameworks has been scanned.

The other important subjects of the research are historically survived (in Appendix A), because their literature reviews are basics and principles of DSP text books. They are: filter-bank, speech separation and NMF. The short surveys are appended after Chapter 6.

2.2 Literature Review of Overlapped-Speech Detection

Speaker diarization splits speech signals of several speakers from their dialogue conversation. The input signal is a conversation where only one speaker is speaking while the other speakers are silent. When that speaker quiets, one silent speaker continues the speaking, and so on. This format of dialogue talking is identical. The spontaneous conversation contains durations of an overlapped-speech of multi-speakers. Any duration of overlapped-speech can be neglected if it does not have regarded information. Sometimes, these durations have regarded information. In this case, the overlapped-speech signal is segregated to its original speakers' signals. The neglecting and the regarding of that information depends on details and contents of the conversation.

In the second half of the 1980s, the speaker diarization field began and was pioneered by several researchers. *O Ghitza* discussed the title of the overlapped-speech by analyzing and synthesizing the overlapped-speech. In 1986, he matched the synthesized speech with its representation [31]. In 1987, he tried to solve the overlapped-speech problem by the auditory nerve representation [32]. In 1992, *Furui* investigated the recognition of the speech under the overlap-speech situation [33]. In 1996, *Kobyashi, Kajita, Takeda and Itakura* extracted the audio features of the overlapped-speech signals [34]. The first attempts for separating an overlapped-speech segments were in 1997 and 1998 by Taniguchi, Kajita, Takeda and Itakura [35].

In the new millennium (2000), achievements of the speaker diarization and the speech separation of an overlapped-speech conversation, are three main scientific activities: post-graduation thesis and dissertations, published articles (papers, presentations and reports) and applied speaker diarization corpuses.

The following are the thesis and the dissertation which are published for the related research area (the remarked PhD thesis and dissertation about the speaker diarization and the overlapped-speech detection are limited number):

About his research, *Xavier Anguera Mir'o* wrote his thesis in 2006. The thesis contains the most important subjects those related to the speaker diarization area. He nick-named the speaker diarization as "who spoke when?". The research thesis covers different types of applications, e.g. the broadcasting [36]. The achievements of Xavier are remarkable among the other achievements in this field. About that field, other papers of Xavier are presented later.

In 2008, thesis of *Kofi Agyeman Boakye* focuses on the audio segmentation, which is the first main phase of the speaker diarization process. In addition to the diarization of the dialogue speech,

chapter of the thesis is a research about the overlapped-speech detection. The achievements of Boakye are remarkable among the other achievements in this field [37]. About that field, other papers of Boakye are presented later.

In 2008, thesis of *Scott Otterson* focuses on the using of the locations of the conversation-speakers for the speaker diarization process. The thesis, also contains a chapter for the overlapped-speech detection. The researcher used the existing application of the NIST corpuses for the meeting speaker diarization [38].

In 2009, the application of *Eugene Koh Chin Wei* thesis is the broadcasting of the news and the recording of the meets [39]. The researcher used the existing applications Hub4 corpuses for the news-broadcasting.

Hornero proposed an overlap labeling method as an integrated component of the *Viterbi*-decoding algorithm for the clustering phase of the speaker diarization [40]. The thesis presents helpful and useful algorithms for that field.

Stephen Shum (2011) used the common algorithms for implanting the speaker recognition and the speaker diarization. The researcher used the applications of the multilingual CallHome (a corpus of multi-speaker phone conversation). The researcher uses the NIST2000 for the speaker recognition [41].

Thesis of *David I-Chung Wang* (2012) details the two main parts of the speaker diarization: the speaker segmentation and the speaker clustering [42].

Jordi Luque Serrano (2011) investigated the environments of multi-sensors for the tracking and the diarization of speakers. His thesis presents the common relationships among the speaker recognition, speaker diarization, and tracking identification and tracking verification [43].

In her thesis, *Mary Tai Knox* (2013) presented the limitation of the speaker diarization and the future directions for that [44]. The achievements of *Mary* are remarkable among the other achievements in this field. About that field, other papers of *Mary* are presented later.

In 2013, *Nguyen Trung Hieu* thesis introduces the two phases of the speaker diarization with the application on the meeting-domain [45].

In addition to his rich publication on the speaker diarization area, thesis of *Simon Bozonnet* (2014) supported his published paper. In addition to the linguistic diarization, he presented the hierarchical clustering scenarios for the speaker clustering [46]. The achievements of *Bozonnet* are remarkable among the other achievements in this field. The other papers of *Bozonnet* are presented later.

Hector Delgado Flores (2015) detailed the binary-key modeling for the overlapped-speech detection and for the cross-session speaker diarization process [47].

In 2015, *Sree Harsha Yella* thesis focuses on the speaker diarization of the spontaneous conversation on the meeting-room [48].

The second activities are the dedicating corpus applications for the speaker diarization and the overlapped-speech detection. Design and implementation of those corpuses, almost used the latest researches in this field. The following are the famous software applications to perform the above tasks:

In 1998, *Zhao, Wegmann and Gillick* [49], then in 1999, *Guo, Zhu and Shi* Introduced the application CU-HTK MANDARIN [50]. It is the first broadcasting-news transcription system. The University of Hertfordshire submitted the Theorizing Visual Art and Design (TVAD) simulation software [51]. Also, the *Göteborg University* produced the MultiTool simulation software [52]. A *pyAudioAnalysis* is An Open-Source Python Library for Audio Signal Analysis [53]. The Rich-Transcription series are RT02, RT03, RT04, RT05, RT06, RT07 and RT09. *C Barras, E Geoffrois, Z Wu and M Liberman* [54] presented "Transcriber", a tool for assisting in the creation of speech corpora, and describe some aspects of its development and use. Transcriber is designed for the manual segmentation and transcription of long duration broadcast-news recordings. The EDI corpus is introduced by the MIT and the University of Edinburgh.

Each thesis or application is part of and/or complement to published paper(s). The third activities are those published papers. The following are the most important papers on the overlapped-speech detection and speaker diarization:

K Boakye, et al. [9] paper contains an overlapped-speech detection algorithm for the spontaneous conversation. In [55], they updated that algorithm. The updated algorithm supports the speaker diarization field. The update is by adding the other coefficients and by warping the coefficients which modify the speaker segmentation process. The improvements improve the overall speaker diarization. They noted that the relative Diarization Error Rate (DER) improvements of AMI (diarization meeting corpus) is about 3 times the DER of the overlapped-speech segments in case of the post processing algorithm. In [56], they presented paper to describe how to improve the overlapped-speech handling to aid of the achievement of the speaker diarization process.

L R Dai, et al. published the [57] for the segregation job of the overlapped-speech job. They supposed that the auditory system of the human kind has an ability for distinguishing and

segregating this composite formula of speech. The approach of the research is the periodicity and the harmonic analysis of the speech conversation signal. The experimental results have been submitted in this paper. The main method of the harmonic principles that is used, could be name as the synchronization of speech frames for the segregation process.

K Laskowski and T Schultz in [58] projected on multi-party meeting and conversation. Main target of them is the increase of robustness and accuracy of the detection job of the overlapped-speech. They described an algorithm for multi-channel overlapped-speech detection. They used a parameters-estimation of the overlapped-speech state. The estimation is unsupervised ML through decoding the combination states of the observation signal of that multi-party meeting and conversation. For cross-checking, the researchers used the well-knowing Rich Transcript (RT) corpus to investigate their algorithm. The relative error reduction of the overlapped-speech detection is about 0.36 per-unit compared with recent achievements.

O Ben-Harush, et al. in [59] presented the co-channel detection of the overlapped-speech and the speech separation of their signals. For the detection of the overlapped-speech segments, the research utilizes the entropy analysis of the processed speech signal. The GMM model is used for identifying the segments of the overlapped-speech. The researchers supposed that the classification phase is divided into two classes: the mixture speech class (the overlapped-speech) and the dialogue speech (the individual speech). For this classification, the border between those two classes is chosen by the conversation hard-threshold. The application which is used for the demonstration and the evaluating is the LDC-CALLHOME American English corpus. The resulting detection is 0.6 per-unit of all the overlapped-speech segments with 0.05 per-unit False-Alarm-Rate E_{FA} (**Chapter 1/1.10 Subjective Test versus Objective Test**).

V Rozgic, et al. [60] presents a multimodal algorithm for the speaker diarization process. They used two hardware components for building the HMM model that represents the problem system. The components are an array of microphones to specify the locations of the speakers, and group of cameras to specify the locations of the participants. In addition to those components, speaker identification algorithm is used to enhance the job. To manipulate the overlapped-speech detection, a likelihood-model for the microphones array observation is used. The researchers modified the concept of the Steered Power Response Generalized Cross Correlation Phase Transform (SPR-GCC-PHAT). They used two methods for the joint estimation, which they are the forward Bayesian filter and the Viterbi algorithm. Their results are an efficient speaker segmentation 0.27 per-unit of

the overlapped-speech for more than 0.94 per-unit on the F-measure scale. In [61], they studied the stereo recording situation by the close-talk microphone for the detection of the overlapped-speech. They suggested a coefficient-derived algorithm by using the spectral similarity between two channels. They note that the similarity is increasing during the period of the dialogue speech format, and decreasing during the mixture overlapped-speech. They concluded that the increasing is because the cross-talk presence. Their suggestion is helpful to distinguish between the dialogue and the mixture formats. By using the DYADIC interaction corpus, the improvement is about 0.26 per-unit.

H Pericás and *F Javier* study [62] focuses on the meeting speaker diarization. They proposed an overlapped-detection algorithm which is supported by the spectral features and the spatial features. The spatial features are taken by using the pairs of microphones then by using the Principal Component Analysis (PCA) method. After the detection of the overlapped-speech segments, the output is applied for the speaker diarization process. This step is for increasing the efficiency of the overall-system by recovering the missing speech. The clustering labels are assigned as multi-labels classification. The comparison between the speaker diarization process and the overlapped-speech detection algorithm, denotes that there are clear distinct behavior of overlapped-exclusion and labeling.

R yokoyama, et al. [63] used the waring lapel microphones for the speech recognition and for the meeting speaker diarization jobs. The audio coefficients are the input of the overlapped-speech detector that is based on the GMM model. They used the Cosine-Correlation-Coefficients for extracting the spectral components and the segments of the speech. By using the meeting corpus, they find that 0.74 per unit better than the traditional recent methods of the overlapped-speech detection and speaker diarization.

W Li, et al. [64] used the fractional dimension coefficients for detecting the overlapped-speech during a spontaneous conversation. They deduced that the chaos degree of the dialogue speech is less than the chaos degree of the mixture overlapped-speech. This deduction indicates that the fractional dimension of the features could be exploited for decrementing between the dialogue segments and the mixture overlapped-speech segments. The comparison between their method with the traditional recent methods shows that there is about 0.81 per-unit improvement.

To exploit the attributes of the overlapped-speech detection for the multichannel semi-blind speech separation, *J M'alek, et al.* [65] proposed semi-supervised speech separation method of stereo-

recorded multi-speech sources. The way uses cancellation filters for computing the potential fixed position of the speakers. The filter computations are chosen after the cross-talk detection process. The former speech sources are separated by the adaptation of suppression technique. The filters are chosen by using the ICA analysis. They tested their overall system by using the exist SiSEC data. The speakers are fixed sometimes and moving other times.

For the same goal, *A Z Wang et al.* [66] proposed noise robust of Underdetermined Blind Source Separation UBSS algorithm. The observation speech signal is an overlapped-speech conversation in addition to a residual cross-talk suppression scheme. The process of the scheme is inside the Short Time Fourier Transform STFT-domain. The algorithm estimates the overlapped-speech array, and then estimates the recovered speech by removing the cross-talk. They suggested a method that use PCA analysis inside that STFT-domain of the mixture overlapped-speech. The GMM model is used for mitigating job against the cross-talk of the recovered speech signals. According to the overall objective tests, the archived results are improved.

S Shum, et al. proposed an approach to speaker diarization based on the Total-Variability to speaker verification [67]. The work is done in applying factor analysis priors to the speaker diarization problem. They arrived at a simplified approach that exploits intra-conversation variability in the Total-Variability space using Principal Component Analysis (PCA). By their proposed methods, the experiments achieve modified performance of 0.9% DER for the speaker diarization. The achievements are done for the added signals of the telephone lines data. The used corpus for those simulations is the NIST-2008-SRE.

P Kenny, et al. reported their achievement for the speaker diarization of the spontaneous conversations via telephone channel. In Johns Hopkins University, they started the title recognition of the robust-speaker [68]. They developed a speaker diarization systems, and they conducted experiments by the corpus NIST2008 for adding the data of the telephone channel. Their scheme consists of: a bottom-up clustering scenario, a point detection algorithm to support the scenario, regular algorithms to cluster the speakers and a different BIC algorithm. The different BIC algorithm is called Variational Bayes System (VBS). The system uses different speaker-factor in the Speaker Recognition (SR). The system modified the DER objective tests by 1.0% for that telephone channel. The overall decreasing of the DER is about 0.85 Per-Unit. That reduction is achieved from the traditional bottom-up scenarios. In contrast with the tradition Clustering, they used the soft-masking instead of the binary masking. Their bottom-up scenario decreases the DER

objective tests by: 0.35 PU & 0.50 PU for the traditional algorithms.

H A Kadhim, et al. in [13, 14] presented a novel method that improvises the algorithm for supervised speaker diarization. The algorithm supposes that the database of the speakers is available. Initially, the database and input observation signal of the speakers, are prepared. The audio features are extracted from the database and the observation signal. Instead of the using of one of Mel-Frequency Cepstral Coefficient, Perceptual Linear Prediction, or Power Normalized Cepstral Coefficients, (independent concatenating features) combination of all of them have been used. The combination form of these features is independent, i.e. they are concatenated in the feature matrix. The comparison between features of observation signal and statistical properties of database features, is made. The comparing algorithm is used to make the decision of the logical mask of the comparison. Both of bottom-up and top-down scenarios collaborate to complete the last decisions successfully. Diarization Error Rate test denotes that combination of features is than errors than any one alone.

Marie-José Caraty and Claude Montacié investigated phenomena, such as the causes, the effects, and the handling of interpersonal or intergroup conflicts [69]. Data on the social and psycho-binary phenomena are collected from people who are involved in the conflict, witnesses of the conflict, or, by extension, looking at a recording of the conflict escalation between the protagonists. From the recordings, they extracted A large quantity of audio and/or video metadata, such as the conversation, the face, and the gesture interactions. The conversational interactions during political debates have been studied to develop an automatic conflict detector from voice analysis. A reliable detector of conflict would be useful for many applications, such as security in public places, the quality of the customer services, and the deployment of the intelligent agents. The development of such a system requires modeling of the conversational interactions as well as the search for specific interactions in relation to a given measure of conflict handling.

H A Kadhim, et al. [70-72] algorithm is a single channel 2-speaker. During the spontaneous conversation, there are times when the speakers are speaking simultaneously (overlapped-speech, or mixture format). Other times, one speaker is speaking and second speaker is silent (dialogue format). Output speech of the algorithm are the segregated dialogue and mixture. At first, the algorithm extracts the RASTA-filtered Perceptual Linear Predictive (PLP) audio features of the conversation. Using the k-means algorithm, the features are clustered into three labels: first speaker, second speaker, and mixture. To find the switching instants from any format to adjacent

another format, the novel concept “Groups” is contributed. Fundamental-group is a collection of the 0.1-second adjacent labels. The variance value of the PDF of each group is the indicator of the speech format: dialogue (low variance) or mixture (high variance). To adopt the useful features and avoid the harmful features, pattern-recognition principals are exploited to optimize the goodness of the groups. Good group contains useful features and bad group contains harmful features. Hierarchical scenarios support the optimized re-clustering by a feedback loop. According to subjective tests, the algorithm performance is excellent. For 300 experimented conversations, average E_{MISS} , E_{FA} and E_{OVL} testes of the outputs are: 0.4%, 1.9% and 1.0%, respectively (**Chapter 1/1.10 Subjective Test versus Objective Test**).

2.3 Literature Review of Speech Separation by Non-negative Matrix Factorization

Non-negative Matrix Factorization is a mathematical method which is efficiently exploited to reduce the matrix dimensionality. Historically, in beginning of the 1970s, Non-negative Matrix Factorization NMF was called Self-Modeling Curve Resolution SMCR [73]. In the middle of that decade, the terms Factorization and Non-negative-matrix appeared on literatures of several mathematical researchers, e.g. *A Berman*, *R J Plemmons* and *L B Thomas*. First attempts for factorization of the non-Negative matrix started during the 1980s by *S L Campbell*, *G D Poole*, *R D Paola*, *J P Bazin*, *F Aubry*, *A Aurengo*, *F Cavailloles*, *J Y Herry*, *E Kahn*, *J S Kahn*, *J Shen* and *G W Israël*.

In the 1990s, early work on the NMF was performed by mathematical researchers. They are the Finnish group: *P Paatero*, *U Tapper*, *A Berman*, *R J Plemmons*, *P Anttila*, and *O Järvinen*. In that decade, *Daniel D Lee* and *H Sebastian Seung* published their papers of NMF, then the term became well-known [74, 75].

After the introduction of *Lee* and *Seung* well-known articles [74, 75], the decomposition ability of the NMF is for-and-only-for the positive-element matrices. Applying of that ability has been started several years later. The first attempt, of that application was by *Hoyer* on the Non-negative Spares Coding [76]. He gave a simple efficient multiplicative algorithm to find optimal values of the hidden components. He illustrated how the Basis Vector can be learned from the observed data. His effective proposal method is simulated and demonstrated. On the same field (the Spares Coding), *Eggert* and *Edgar Korner* showed how to merge concepts of the Non-negative Factorization with sparsity conditions [77]. It is a multiplicative algorithm which is comparable in

efficiency to the standard NMF. That can be used to gain sensible solutions in the over-complete cases. The case for learning and modelling the arrays of receptive fields arranged in a Visual Processing Map, where an over-complete representation is unavoidable.

In his literatures, *Schmidt* presented first single channel speech separation by using the NMF [78, 79]. The separation is performed in a low dimensional feature space which is optimized to represent mixture speech. For each speaker basis, the over completed is done by the term “Sparse Non-Negative Matrix Factorization”. The mixture speech of two speakers is separated by mapping the mixture onto joint bases of them. The method is evaluated in terms of word recognition rate, on the speech separation challenge data set.

For MIREX2007-project [80], *Vincent*, *Bertin* and *Badeau* presented two NMF methods for polyphonic pitch transcription [81]. The Polyphonic pitch transcription consists of estimating the onset time of each note within a music signal. The estimation includes the duration and pitch of that musical note. According to [81], the NMF appears well-suited to this task since they can provide the true representation of which musical instruments are playing. *Vincent*, *Bertin* and *Badeau* proposed a simple transcription method using: minimum residual loudness NMF, harmonic comb-based pitch identification and threshold-based onset and offset detection. They investigate second method incorporating harmonicity constraints in NMF model.

For the musical analysis, *Févotte*, *Bertin* and *Durrieu* have applied NMF measurements with the derived algorithmic. The main approaches for their article are: The Gamma-Markov Chain Prior (GMCP) with its inverse (IGMCP) and the *Itakura-Saito* distance [82]. It is possible to define HMM discrete version (Markov-Chain) on inverse Gamma function of the random variable in straight-forward procedure. The relationship of the *Itakura-Saito* distance and the NMF-based cost functions, is considered in this article. The relationship between the traditional Euclidean distance with the *Itakura-Saito* and *Kullback-Leibler* distances. The article contains comparison between these methods. Power Spectrum sequence is used, also in their algorithm. Real-time sequence is considered in the implementation.

For SiSEC2008-Campaign, *Févotte* and *Ozerov* utilized the multichannel NMF in convolutive mixtures for audio separation [83]. The NMF-based separation is supported by the *Itakura-Saito* distance. The distance is modelled by the statistical Gaussian elements. Using the following two methods, they addressed the estimation of the speech mixture and the targeted-speech: 1) By the likelihood maximizing of the multi-channel using the expectation and the maximization algorithm.

2) By the likelihood maximizing of the individual model of the overall-channel. This method based on the NMF of the multiplicative algorithm. For linear approximation, Short-Time-Fourier-Transform (STFT) is applied on an instantaneous mixture of audio and speech signals.

For NMF-based audio, sound and speech separation, *F'evotte, King and Smaragdis* suggested an optimization method. The method exploits two algorithms: 1) magnitude power 2) cost function [84]. They looked at the two applications: 1) single channel audio, sound and speech separation. 2) Recover the missing audio (e.g. music) by the interpolation mathematical approach. They optimized the audio and speech parameters in their simulations. As well as, they discussed the influence of those audio and speech parameters on their simulations.

For audio separation, *Virtanen, Cemgil and Godsill* used Bayesian extensions to NMF [85]. A conjugate Gamma chain prior enables modelling the spectral smoothness of natural sounds in general. Other prior knowledge about the spectra of the sounds can be used without resorting to too restrictive techniques where some of the parameters are fixed. The resulting algorithm, while retaining the attractive features of standard NMF such as: fast convergence and easy implementation, outperforms existing NMF strategies in a single channel audio source separation, and detection task.

W Wang compared between Instantaneous and Convolutional NMF in [86]. The comparison includes the models, the algorithms and the applications to audio pattern separation of them. The convolutional NMF model, which has an advantage of revealing the temporal structure possessed by many signals, has been proposed. *Wang* provided a brief overview of the models and the algorithms for both the instantaneous and the convolutional NMF. He focused on the theoretical analysis and the performance evaluation of the convolutional NMF, and their applications to Audio Pattern Separation.

HA Kadhim, et al. [22, 87] proposes single-channel blind speech separation algorithm. Input signal is 2-speaker section (segment) of mixture speech. The section is output of overlapped-speech detection process. The algorithm consists of four sequential techniques: filter-bank analysis, Non-negative Matrix Factorization (NMF), speaker clustering and filter-bank synthesis. At first, the mixture speech signal passes through filter-bank analysis, to produce 65 signals. NMF factorizes spectrogram of each signal into 24 sub-signals. The 1560 (65×24) sub-signals are separated into two speakers. The separation is performed without identifying the speakers. To identify the speaker of the separated sub-signals, speaker clustering is exploited. Since the speaker clustering needs

speaker segmentation, standard framing is used to partition each sub-signal. Filter-bank synthesis sums the identified partitions, to produce two output separated speech signals. Masking and phase-angle recovering are merged with the clustering. The algorithm is simulated and then tested for 51 conversations (including TIMIT-library speakers). The average SAR, SDR and SIR objective tests are: 5.06dB, 3.75dB and 12.47dB respectively.

In his Master's thesis, *Cauchi* presented applications of the NMF for the Auditory Scenes Classification [88]. He noticed that the short artificial examples which consider the non-stationarity of the spectral content of the sound sources, improves the source detection. Classification method is applied to a corpus of soundscapes of train stations, and the results are compared with previous classifications methods. He concluded that the NMF significantly improves the classification.

Zheng, et al. classified the NMF matrices using the similarity between theses matrices for the source separation of the speech mixture signal with the music [89]. Signal-to-Noise thresholds are found experimentally. The speech signal and the recovered music are recovered using the NMF matrices. Their simulations outputs illustrate that the separation of these speech and audio signal are efficient. That efficiency is concluded by the comparison with the well-known methods of audio, sound and speech separation.

For noise-robust automatic recognition of the speech (SR), the [90] authors utilized the NMF methodology for the audio and speech separation. In the article, the audio coefficients have been extracted by a filter-function using the 2D-*Gabor*. The function extracts these features of the speech and audio signals by the Time-Frequency representation and the spectro-temporal. The [90] improves the "back-end classification". For instance, the article modified the recognition rate by 0.5 Per-Unit. The input Signal-to-Noise Ratio is -6 dB.

For cluster frequency basis functions of monaural sound source separation, *Jaiswal* detailed NMF-based algorithms in his PhD thesis [91]. He attempted to solve the problem of clustering NMF-basis functions. He used shifted NMF as a method of clustering the basis functions obtained via standard NMF. The shifted NMF clustering method aims to cluster the frequency basis functions obtained via standard NMF to their respective sources. It is by making use of shift invariance in a log-frequency domain. He improved the separation performance of the standard NMF algorithm obtained through use of an improved inverse Constant Q-Transform.

Gao, et al. proposed un-supervised audio separation method. The method uses filter-bank technique. The Filter-bank is configured as gamma-tone array of filters. In addition to the filter-

bank, audio coefficients have been extracted for the processed audio signals [92]. The audio mixture signal is separated by this configuration of the filter-bank rather than the tradition uniform configuration of the filter-bank. They presented the audio separation method by the NMF-based Itakura–Saito distance. The Quasi Estimation Multiple framework is used for their derivation. In [93], their approach is 2-Dimensional Non-Negative Matrix Factorization. The NMF-base approach is modified for the maximum a posteriori probability Variational method. The method generalizes criterion for variable sparseness, and incorporates into the basis features. The method is computationally efficient, and demonstrated on extracting features from image and source separation. The basis features of an information bearing matrix are extracted by regularized priors. In [94], their method is adaptive sparsity non-negative matrix factorization. The NMF decomposes the information-bearing matrix into 2-dimensional convolution matrices representing the spectral dictionary and temporal codes. They derived optimization uses the Variational Bayesian to find the parametric-sparsity NMF. Variational Bayesian is a technique to approximate the intractable integrals arising in Bayesian Interference and Machine Learning fields. The separation experiments process the mixture of an audio signals of single channel. In addition to that, they used the spectrum of the dictionary under the sparsity conditions. The output resulting separated signals denote that the method is efficient by the comparison with the tradition methods.

2.4 Literature Review of Informed Speech Separation

Informed speech (source) separation is supervised or semi-supervised machine learning DSP system. There are several methods and approaches to perform that job. The review surveys the most important themes. The last review focuses on the NMF-based informed speech separation, which is in touch with the Chapter 5 algorithm. The recent themes within the informed source separation framework are:

2.4.1 Video-Assisted Source Separation

For Audio-Visual Dictionary Learning (AVDL), *Q Liu, et al.* proposed new method in [95]. Their titles are linked with that AVDL technique. The cross-modality differences are the main titles. Their challenge is the differences of the cross-modality of the system. These are: difference of the dimensions, difference of the sizes, difference of the scales and the complex of the computation. They used coding-learning method by developing the AVDL technique. They used adaptive

dictionary for the above differences of the scale and the size. They considered the indexing of the fast-search concept. The cross-modality, also is considered. Their suggested method is linked with speech separation application. The mimicking aspects are considered also. They derived Audio-Visual Speech Separation method. The proposed method is AVDL interacts with the Mandel speech separation method. The Mandel separation method is an audio Time-Frequency masking method. Using the subjective tests, they assigned those methods. They compared these methods with the traditional methods. The well-known corpus (LILiR-Twotalk) is used for that comparison. The authors with *S M Naqvi* [96], design a scheme to reject the non-stationary audio coefficients. The designed scheme modifies the robustness of their model. The scheme reduces the complexity of the computation. The coherency with the mathematical interpolation configure these coefficients. These configuration is used for the source separation job. That job is performed on the frequency-domain. To cross-check the proposed algorithm, multimodal data are used. Simulation outputs show the performance improvement with the proposed algorithm.

A Kazemi, et al. exploited the relationship of the speech signal with the movements of the lips. The lips movements are monitored by the recorded video signals. These relationship modified the Audio-Visual Speech Separation (AVSS) methods [97]. The link between the audio and the video could be modeled by the neural interaction. The parametric prediction of the lips movements is modeled by the “audio-observation”. The authors contribute an objective Mean-Square-Error MSE experiments. The Errors is the difference of that estimated parameters with the targeted-speech parameters. The MSE function is optimized to perform that estimation. The authors contributed rules by the audio-visual simultaneously action. The statistical kurtosis is used for the modeling measure. The article presented the simulation output results. Those results are compared with traditional speech detection methods. Signal-to-Interference Ratio (SIR) objective test is used to evaluate the performance of the article. According to the compression with the ICA-based and AVSS-based speech separation, the proposal is efficient.

2.4.2 Spatial Audio Object Coding

A Ozerov, A Liutkus, R Badeau and *G Richard* introduced Coding-based Informed Speech Separation (CISS). They related the general Informed Speech Separation with the statistical Source-Coding theorems [98]. The Coding-based ISS encodes the speech source signals and the input mixture speech signals, as well. The proposed method is more efficient, compared with the

tradition methods, because: 1) The required bandwidth is available like the source coding case. 2) The transmitting bit-rate is similar for the conventional ISS. Mathematically, their Coding-based ISS is done by the Non-negative Tensor Factorization (NTF). The Coding-based ISS Rate-Distortion of their method is better than other traditional ISS methods.

By exploiting the technique of sparse coding, *L Zhen, et al.* proposed an effective approach to discover several 1-Dimensional of the time-frequency spectrogram matrices of the input mixture speech signals [99]. They showed that those 1-Dimensional are related to their spectrograms transformation points. By the using of the collections of those matrices in the hierarchical (bottom-up and top-down) scenarios, they approximated the mixture-speech spectrogram matrix. The targeted-speech signals are obtained by the Least Squares (LS) algorithm. The approximated estimation of the matrix which attained by the article proposal, is efficient compared with the traditional speech separation methods. The demonstration outputs appear that the method improve the recent methods.

In [100], *W Nogueira, et al.* investigated the situation of mixture speech signal with other sound signals. The Deep Recurrent Neural Network (DRNN) is used for that situation. The cochlear model of the system is the approach for the used DRNN. Optimization process is exploited for the masking task. The approach is tested by a male HSM sentence. The sentence is combined with female speech. The goal of the test is the source separation. The used DRNN have 2 levels. The tests include eight traditional listeners. The two DRNN levels provide clear modifications for the output speech intelligibility. Those modifications are attained by the Vcoded-speech experiments.

2.4.3 Reverberant Models for Source Separation

S Arberet and P Vanderghenst in [101] they showed that the performance can be improved by the using of low-rank of the targeted-speech spectrogram matrices. The article algorithm estimates the targeted-speech by the analysis of the sparse and by the low-rank matrices of the targeted-speech spectrograms. The tests of input musical observation signals present an improvement of the output separated signals. The Signal-to-Distortion Ratio (SDR) improvement is about 2.0 dB.

S Leglaive, et al. in [102] presented probabilistic mixing filter to assist the job of the source separation. The reverberant recording constrains are regarded for that. They proposed specific R-order propagation. The processing is in the frequency domain. Its direction is coming from the targeted-sources signals and is going to the mixture signals. The process is autoregressive in that

domain. The processing is a prior to derive a Maximum A-Posteriori (MAP). Expectation-Maximization (EM) model supports that process. Demonstration results are better compared with the traditional Maximum Likelihood (ML) methods.

A Asaei, et al. in [103] speakers model is specific 2-D plain. The speech paths of those speakers are multi. The recorded the targeted-speech and audio signal is modeled by the Room-Acoustic. The dimensions of that room describe the locations of the speakers. Approximately, the speech sources are virtualized as images of the sources speakers. The clustering process is attained for those images. That process identifies that room acoustic response and its dimensions. The acoustic reflection and absorption are considered also. That system parameters are extracted from the description. The simulation of that acoustical system is realized for the recording.

2.4.4 Score-Informed Source Separation

S Ewert, et al. in [104] article, describes the modification for the Score-Informed Source Separation title. They compared between different strategies of that informed source separation. They focus on the musical separation methodology. Main applications of that musical signal separation are the mixing of the stereo and the surrounding. DJ remixing and the musical instruments level equalization are regarded also. The correlation between the original musical signal with surrounded signals are used for that processing. That guided separation is achieved for the musical instrumentation. The separated sounds are high quality and have clear robustness.

J Driedger, et al. in [105] provided a decomposing musical system. They provided a method to deploy that scored-information for the audio signal sources. They arbitrary select, then analyze the notes by the improvement of source recording. The system contributes a modified method for the audio application such as the editing. It has the good ability to test the audio and the source separation methods.

2.4.5 Language-Informed Speech Separation

Z Q Wang, et al. in [106] contributed a Phoneme-Specific Speech Separation (PSSS) algorithm. Instead of the phoneme training one-by-one, they trained multi-phoneme. Since each multi-phoneme cover specific frame, the required processing is done for that corresponding frame of those multi-phoneme. The traditional Automatic Speech Recognition (ASR) is exploited for the identifying of each phoneme frame. ASR methods provides the language-model which support the

informed speech separation. That phoneme model has good merit against the small changing. The experiments confirm that the recognition job is supported by this process. The CHiME corpus for the speech separation is used in the simulation tests. The system has been tested objectively to check the intelligibility of the output separated speech. It has been used for checking the ARS achievements, also.

2.4.6 User-Guided Source Separation

R Hennequin, et al. in [107] proposed underdetermined method for the audio separation. The method exploits the signals of the spoken references and/or the signals of the sung reference. These signals could change the blind audio separation to informed audio separation process. The method describes the difference- model of the targeted-audio signals, and the spoken- and/or the sung-references signals. The method uses the pitch for the difference-model and the time-lag of these signals. The proposed method is a novel contributed method.

Q Wang, et al. in [108] proposed significant approach to separate the mixture speech signal. The mixture speech is single channel. The main idea of the separation uses the user exemplar to assist the process. The added exemplar speech is a related recorded speech. The training process of that method is speaker-dependent data which has specific relationship with the conversation subject(s). The added exemplar speech should be utterance-dependent also. The pattern of that utterance is done similarly to the original pattern of the mixture conversation. According to the article, that method does not need speaker dependent data, and yet exceeds the performance of traditional separation models for separate the male and the male mixture speech signals. After the testing and the comparing with recent method, the method is efficient.

2.4.7 Dictionary-Based Methods

In his PhD thesis, *Lefevre* presented dictionary learning methods for single-channel audio source separation [109]. He provided a group-sparsity inducing penalty specially tailored for Itakura-Saito NMF. He presented an online algorithm for Itakura-Saito NMF that allows learning dictionaries on very large audio tracks. The memory complexity of a batch implementation NMF grows linearly with the length of the recordings. It becomes prohibitive for signals longer than an hour. In contrast, his online algorithm can learn NMF on arbitrarily long signals with limited memory usage. In short mixed signals, the blind learning becomes very hard and the sparsity do not retrieve interpretable

dictionaries. He described an extension of NMF to consider time-frequency localized information on the absence/ presence of each speech/ audio source. He also investigated inferring the information with tools from machine learning.

2.4.8 NMF-Based Informed Speech Separation

M Fakhry, et al. in [110] addressed an underdetermined audio separation method. The method is assisted by the existing data for the sources. Local GMM model for the mixture signals are used to overcome the main separation problem. For the mixture signals, in addition to the GMM the covariance-matrix are used to extract the proper parametric matrices. To extract those parameters, iterative algorithm is suggested in this article. The Maximum-Likelihood (ML) is used to re-configure the parametric matrices by the exploiting of the already exist data. Non-negative Matrix factorization (NMF) is applied for the factorization/separation process. The variance matrices of the virtual source signals are represented by the multiplication of the 2 factorization matrices: The Spectral-Basis and the Time-Varying Features matrices. The first matrices of the virtual speech (or source) signals has been trained. Other non-useful (corrupted) data are neglected in that training process. Those trained data assist the separation process by move their amplitude. This paper method is experimented by the simulation of real-time mixture signals. The outputs denote that the method is efficient.

Tu, Jiao and Xie proposed a semi-supervised machine-learning system for the speech and audio mixture signals separation. The observation mixture signals are recorded then process by the NMF technique [111]. The Sparseness NMF (SNMF) separation process to separate the mixture speech and then music signals one-by-one. The experiments to investigate that system include 10 Mandarin-library of the speech. The objective Speech-Music Ratios (SMR) tests checked the 10-speaker observation input speech and music signals. That semi-supervised ML audio and speech separation system does not have specific constraints. The tests appear that the system improves that separation process significantly.

Joder, et al. have comparative-review about the SNMF-based (speech) source separation. The NMF-based separation systems are supervised ML system and/or semi-supervised ML system [112]. The compared reviews are according to the Wiener-Entropy rules. The objective testing evaluation involve the spontaneous conversations with the music. Those results denote that the enforcing of the SNMF conditions. The training phase provides best results for the supervised

NMF.

Roux, et al. proposed the deep Non-negative Matrix Factorization and the deep architecture of the Non-negative network. That network is architecture by the unfolding of the NMF-based optimization loop [113]. The network architecture is trained to find the optimal source separation process. The optimization is attained to find best parameter-matrices of the NMF. By the neglecting of the constrains, back-propagation is the algorithm for that optimization.

For Deep Neural Network (DNN)-based supervised speech separation, [114] authors exploited spectro-temporal structures using NMF. With non-negative constrains, NMF can capture the basis spectra patterns. As the basis functions, they added to the original output of DNN to reconstruct the magnitude spectrograms of speech and noise with the non-negative linear combination. Reconstructed spectrograms, a discriminative training objective and a joint optimization framework are used for the proposed model. Systematic experiments show that the proposed model is competitive with the previous methods in monaural speech separation tasks.

Bouvier, et al. used a source/filter technique to speech separation by the NMF [115]. The NMF technique has its adaptive-constraints. The constraints are the baseline of the source/filter technique to perform speech separation algorithm by the weighing of them through that process. That source/filter technique is a semi-supervised machine-learning system. For the data training, the filter-basis is approximated inside the separation process for the speaker- phoneme. That approximation includes the constrains of that source/filter technique. The experiments of that technique appear the increment for the adaptive constraints simulation. The speech separation by that technique is better for the fully-supervised machine-learning system.

In [116], *Gao, et al.* approach solves single-channel source separation problem by resort exclusively to the independence waveform criteria. The separation is supervised system by training the prior information before the separation process. The algorithm is complete and efficient to NMF.

H A Kadhim, et al. in [23, 117], the input signals of the informed speech separation are: 2-speaker mixture section (segment), plus section of individual speech of each speaker. The main input is spontaneous conversation between 2 speakers. The input is already processed by overlapped-speech detection. The detection had isolated the input into two speech formats: dialogue and mixture. The dialogue is when only one speaker is talking. The mixture when the speakers are talking simultaneously. The mixture speech is the observation input signal of the informed

separation. The dialogue had been processed by speaker diarization to produce individual speech section of each speaker. These sections are supposed as virtual-targeted-speech, and their mixture as virtual- mixture. The diarization process assist the separation (semi-supervised machine learning). The training is achieved by homogeneous emerging the virtual-speech with the real-speech, in the NMF. The separation using the NMF technique. The algorithm is simulated and then tested for 341 conversations (including TIMIT-library speakers). The average SAR, SDR and SIR objective tests are: 9.55 dB, 1.12 dB and 2.97 dB respectively. To improve the separation, optimization masking is contributed.

J Fritsch and *M D Plumley* in [118] presented improved algorithm to separate the mixture signal of music and audio. The algorithm extracts the required data from a score of the musical instruments. The algorithm is supervised machine learning system. Basically, the algorithm uses the traditional supervised NMF process. Harmonic constrains are used for the training phase of the system. The MIDI is corrected manually to create the required multi-track database to support that training phase. To compare between this article algorithm and other traditional algorithms, Blind-Speech-Separation BSS-EVAL toolbox with the PEASS toolkit are used together. The evaluations denote that the BSS-EVAL objective tests have clear improvement.

F J Rodriguez-Serrano, et al. in [119] used the Complex Matrix Factorization (CMF) to find the audio parameters. That process needs stable CMF to perform efficient separation process. The shift-invariant CMF has been used to get the NMF ability to CMF. The demonstration of the method denotes for the clear improvements.

N Souviraà-Labastie, et al. in [120] investigated guided system for the musical signals separation. That system is done by the multi-tracking of the recorded signals. An interpretation technique is used for the processed songs. Experimentally, the Joint NMF with a specific transformation are used. The improved Signal-to-Distortion Ratio (SDRI) is about 11.0 dB. That improvement is about 2dB by the comparison with recent articles.

N Guan, et al. in [121] proposed a transductive NMF method (TNMF) to jointly train a dictionary for speech signals of the speakers and the mixture speech signals to be processed. The TNMF trains the descriptive dictionary by encoding the mixture signals more than that trained by the traditional NMF. The experiments on the TIMIT library speech, show that the TNMF-based algorithm better than the ordinary NMF-based algorithms for the speech separation of the mono-phonetic speech mixture signals of specific speakers.

Chapter 3. Overlapped-Speech Detection based-on Stochastic Properties



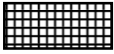
3.1 Introduction

Chapter 3, Chapter 4 and Chapter 5 are the kernel of the research. They contain full description of the contributed algorithm and algorithms. To understand the attributes and the operation of the research, the link between these chapters is very important. The main observation signal of the research (the spontaneous conversation) is the input signal of the Chapter 3 algorithm. The two outputs of the Chapter 3 algorithm are two segregated speech signals. The first output is the dialogue speech signal. This signal is processed by reliable existing speaker diarization toolbox. The toolbox outputs are the isolated independent speech segments of all the speaker, each one alone. The second output of this chapter algorithm is the mixture speech signal (the overlapped speech). It is the input signal of Chapter 4 and Chapter 5 algorithms.

The novel algorithm of this chapter is overlapped-speech detection. The algorithm estimates the time-domain location of the switching instants from format to the another. The input conversation speech signal is framed by the standard overlapping-windowed frames. 13 RASTA-PLP audio features per that frame are extracted. The feature array is clustered into tree clusters: the first speaker (dialogue), the second speaker (dialogue) and the mixture speech of the two speakers. Collection of 0.1-s features are capsulated inside fundamental Group. Sequential variances of the PDF of the groups are calculated, then resides its row of variance array. The variance row is clustered into two clusters: high and low variances. That collection is repeated for 0.2-s, 0.3-s, 3.2-s. The above is repeated for them to produce two arrays: variance and decision (clustering) array. The two patterns of the variances have common area. The area is indication of the recognizing goodness between the patterns. Principles of ML and PR are used to evaluate the relative goodness of these groups. These steps are optimization process to avoid the features inside the worst groups and to adopt the features inside the best groups. The optimization loop elaborates the quality of the features. Re-clustering of the above is efficient to clusters them and detects the switching instants. Hierarchical clustering scenarios improve the last step, the masking.

The algorithm is experimented using arbitrary large number of (female and male) speakers. The input signals are formulated as 2-speaker conversations of the speakers. The Subjective and objective tests indicate that the algorithm is highly efficient. That evaluation is concluded by the comparison of the algorithm standard tests with the recent reliable corpuses tests [70-72].

3.2 Functional Block Diagram and Illustrative Waveforms

Input of the overall system, for the overall thesis is the input of the overlapped-speech detection process. In this chapter, the input is prepared of 5 minutes of a spontaneous conversation. Large number of the prepared conversations, cover all the required possibilities of genders, and the randomly conditions of such conversations. For those reasons, long period (5 minutes) of spontaneous conversations have been prepared. Each conversation consists of 10 periods (segments), each segment is 30 second (s) in length. Suppose that the speakers involving in each conversation, are two speakers: the first is Female (F) and the second is Male (M). The conversation is a dialogue between them, where one of the speakers, individually is talking but another speaker is not talking (is silent). During this conversation, there are frequent patches of overlapped-speech between them, i.e. they are talking together at the same time (called mixture speech FM) [69]. Continuing with chapter 1 description, (a)/Figure 3.1 is a typical sketch of this conversation, where the period of each F, M or FM segment is 30 s. The vertically-lined  segments are the periods of the F dialogue speech, which they are 3 of 30-s segments, i.e. one and a half minutes in total. The horizontally-lined segments  are the periods of the M dialogue speech, which are 3 of 30-s segments. i.e. one and a half minutes in total. The mesh-lined  segments are the periods of the FM mixture speech, which are 4 of 30-s segments, i.e. two minutes in total. As was seen in the chapter 1's: details, block diagrams and sketches, the signal of the dialogue speech format can be processed by the speaker diarization algorithms to isolate the speech of each speaker alone [14, 27, 122, 123]. Also, the signal of the mixture speech format can be processed by the speech separation algorithms to resolve the speech of each speaker alone [124, 125].

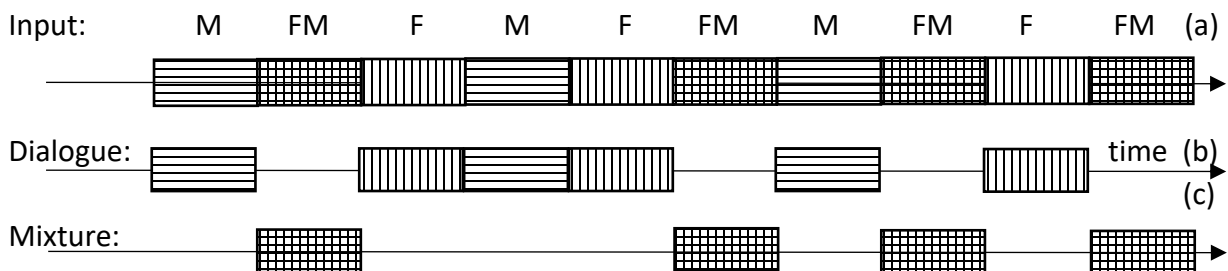


Figure 3.1 Typical sketches of input and outputs of Chapter 3 algorithm. The (a) is the input (i/p) of the algorithm. The (b) is 1st output (o/p) of the algorithm, i/p of speaker diarization (dialogue speech). The (c) is 2nd o/p of the algorithm, i/p of speech separation (mixture speech).

Before the speaker diarization and the speech separation processes, the mixture overlapped-speech should be segregated from the dialogue speech [70-72]. The (b)/Figure 3.1 illustrates the dialogue output signal, and the (c) illustrates the mixture output signal.

Later, for the experiments checking of this chapter algorithm, the TIMIT (form more details, please Ctrl+Click to access its website) standard audio and speech library [20] is used in addition to arbitrary sufficient recorded speech. For this purpose, the F and M conversation in Figure 3.1: a female is talking with a male, a female is talking with a female or a male is talking with a male.

This chapter contains the description of a novel algorithm that achieves this segregation job successfully. Figure 3.2 shows this chapter functional block diagram. Figure 3.1 shows the time domain signal waveforms of its input and its outputs. This chapter block is the first part of the overall system (see Figure 1.5 and Figure 1.6).

3.3 An Algorithm of Overlapped-Speech Detection

This chapter focuses on the first block of the three main blocks of the thesis research overall system (Figure 3.2, Figure 1.5 and Figure 1.6). The block abstracts an algorithm which segregates the input spontaneous speech signal into two output signals: the first output is a dialogue speech signal and the second is a mixture speech signal. The algorithm is a combination of traditional and modified well-known techniques and algorithms. Each step of these combinations has been tested alone, and then have been tested with the other steps (two or more steps). After the primary and the final tests, the best choices are adopted. The primary testing for specific step and the final testing for the overall system is conducted by subjective tests and the reliable objective tests. The sequential algorithm of the algorithm is detailed in the next titles and subtitles of this chapter.

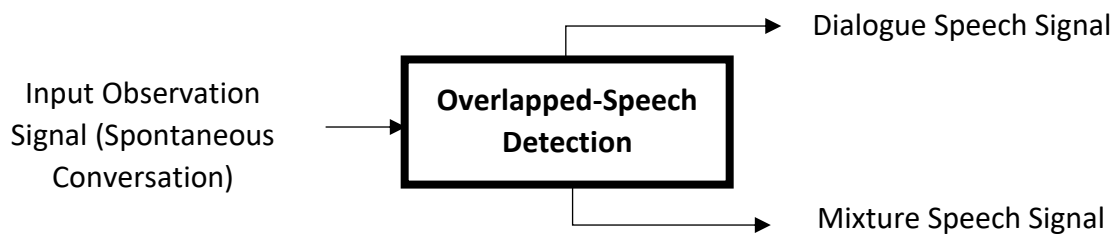


Figure 3.2 General block diagram of the overlapped speech detection process. The Input signal is a spontaneous conversation speech (the (a)/Figure 3.1), i.e. contains a dialogue speech and a mixture speech. The outputs are the segregated dialogue and mixture signals (the (b) and the (c)/Figure 3.1, respectively).

3.3.1 Framing and Overlapping-Window of the Input Signal

According to the reliable references of the audio and the speech-DSP, the optimal period to process any speech signal frame, successfully, is 8 to 20 ms per frame (or its equivalent number of samples). Frequency domain resolution of the range, of such period, is 50 to 125 Hz per sub-band. This range of resolution does not cover the requirements of audio and speech-DSP. The accepted alternative is the 16 to 40 ms period of overlapping-window frames T_w . For this chapter algorithm, $T_w = 32$ ms of the frame with 22 ms of overlapping are chosen, i.e. the hopping period (T_h) is $(32-22=10)$ ms. According to these choices, each frame represents 10 ms of speech signals with negligible 11 ms $(22/2)$ on the previous and 11 ms on the next times of that frame. Either 16000 or 8000 sample/s is the sampling rate of the input signal with a resolution of 16-bit per sample. The audible differences between these sampling rates are negligible as well as the number of bits per sample resolution. The number of input samples per frame N_w is 512 samples for the 16000 sample/s sampling rate (f_s), and 256 samples for the 8000 sample/s sampling rate. The number of the samples for each hopping N_h is 160 for the 16000 sample/s sampling rate, and 80 samples for the 8000 sample/s sampling rate [1, 2, 126].

3.3.2 Extraction of Audio Features by RASTA-PLPC

Audio and speech-DSP describes the audio and the speech signals inside various parametric models. In the time domain, they are represented by waveforms of their amplitude versus the time domain axis. In the frequency domain, they are represented by their sub-bands using Short Time Fourier Transform STFT analysis. For speech and speakers, these time and frequency domains might not have the efficient and enhanced ability for most audio and speech-DSP, e.g. the recognition of different speech or Identification and Verification of the Speakers. The most productive parametric model for the audio and the speech is the extraction of their features/coefficients instead of the time and the frequency domain representations. The well-known algorithms and techniques used to extract the audio and the speech (coefficients) are [127, 128]:

- The Mel-Frequency Cepstral Coefficients (MFCC).
- The Linear-Frequency Cepstral Coefficients (LFCC).
- The Linear Predictive Coding Coefficients (LPCC).
- The RASTA-Perceptual Linear Prediction Coefficients (RASTA-PLPC).
- The Power-Normalised Cepstral Coefficients (PNCC).

In this research chapter, the above techniques have been tested, but after cross-checking of the results among them, the RASTA-PLPC is chosen because it yields the desired efficient features. Linear Predictive Coding (LPC) is an old well-known algorithm that abstracts the time domain sequence $x(n)$ and the frequency domain contents $X(k)$ of speech signal, then utilizes its m coefficients (the a 's) instead of these contents [129]. The LPC is a statistical approach that uses recursion algorithm to predict the current sample $x_p(n)$ by the extrapolation of the m previous samples: $x(n-1)$, $x(n-2)$, $x(n-m)$:

$$x_p(n) = \sum_{i=1}^m a_i x(n-i) \quad (3-1)$$

The resulting current error is the difference between the actual $x(n)$ and the predicted values $x_p(n)$:

$$e(n) = x(n) - \sum_{i=1}^m a_i x(n-i) \quad (3-2)$$

If $e(n)$ is the input and $x(n)$ is the output of equation (3-2), the transfer function of its equivalent system can be represented by a Finite Impulse Response FIR filter. The filter consists of m Delay-Units and m weighting values (the a 's).

The goal of the linear prediction is the estimation of the set of its coefficients $\{a_1, a_2, \dots, a_m\}$ from the m previous samples of the input sequence $x(n)$. For that, the standard solution is by the minimizing of the Mean Square of the Error (MSE):

$$\sum_n e^2(n) = \sum_n \left[x(n) - \sum_{i=1}^m a_i x(n-i) \right]^2 \quad (3-3)$$

To minimising the error:

$$\frac{\partial e_n}{\partial a_j} = 0 \quad (3-4)$$

where,

$$\frac{\partial e_n}{\partial a_j} = -2 \sum_n [(x(n)x(n-j) - \sum_{i=1}^m a_i(x(n)x(n-j))] \quad (3-5)$$

By formulating the equation (3-5) using its equivalent matrix form, then by the applying of the Levinson-Durbin Recursion, the system has become solvable, and the m number of coefficients (the a 's) are extracted [130].

The Perceptual Linear Prediction Coefficients PLPC technique is a modified version of the LPC, by *Hynek Hermansky* in 1989 [131]. Figure 3.3 illustrates the flowchart of first *Hermansky* version of PLP technique. According to that version, the input speech signal passes through:

- The Critical-Band Analysis.
- The Equal-Loudness Pre-Emphasis Filter.
- The Intensity-Loudness Conversion.
- The Inverse Discrete Fourier Transform (IDCT).
- The Autoregressive-Coefficients Solution.

To build the All-Pole Model and extract the required audio coefficients, the above autoregressive solution exploits the Levinson-Durbin Recursion.

According to the Parseval's theorem, the power spectrum $|X(f)|^2$ of $x(t)$ speech signal is the linear scale distribution of this signal, which is the result of the Fourier transformation, then the square of its magnitude. In the linear-scale, each sub-band has equal weight as any another sub-band in the frequency domain. In case of the power (energy) of the speech signal, the distribution is nonlinear. The non-linear distribution is because the fact that the weight of any sub-band is differences in comparison with the other sub-bands. The nonlinearity phenomenon has been adopted through the past 80 years of experimental research. The laboratory experiments focusing on the behaviour of the human hearing, deduced the fact that the major part of the speech power (energy) occupies the sub-bands of the lower frequency range below 1 kHz. The other sub-bands have the minority part of these power (energy). Due to this non-linearity distribution, the concept of the Critical-Band was introduced by *Harvey Fletcher* in 1933 [132]. The laboratory experiments focused on the human ear, especially on the cochlea of the inner ear. There are three recommended critical frequency bands of scaling and distribution: The Equivalent Rectangular Bandwidth *ERB*, the *Bark*-scale and the *Mel*-scale [133]. Figure 3.4 shows the normalized frequency scaling of these critical bands.

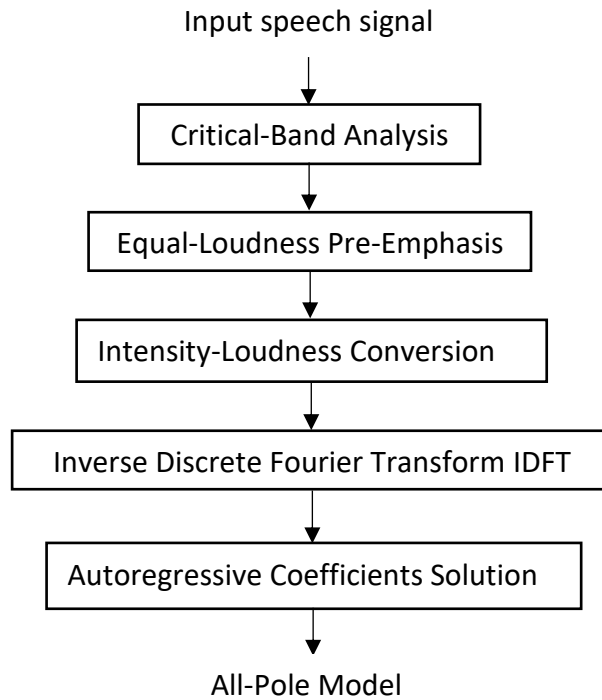


Figure 3.3 Flow chart of the first *Hermansky* version of Perceptual Linear Prediction PLP technique [131].

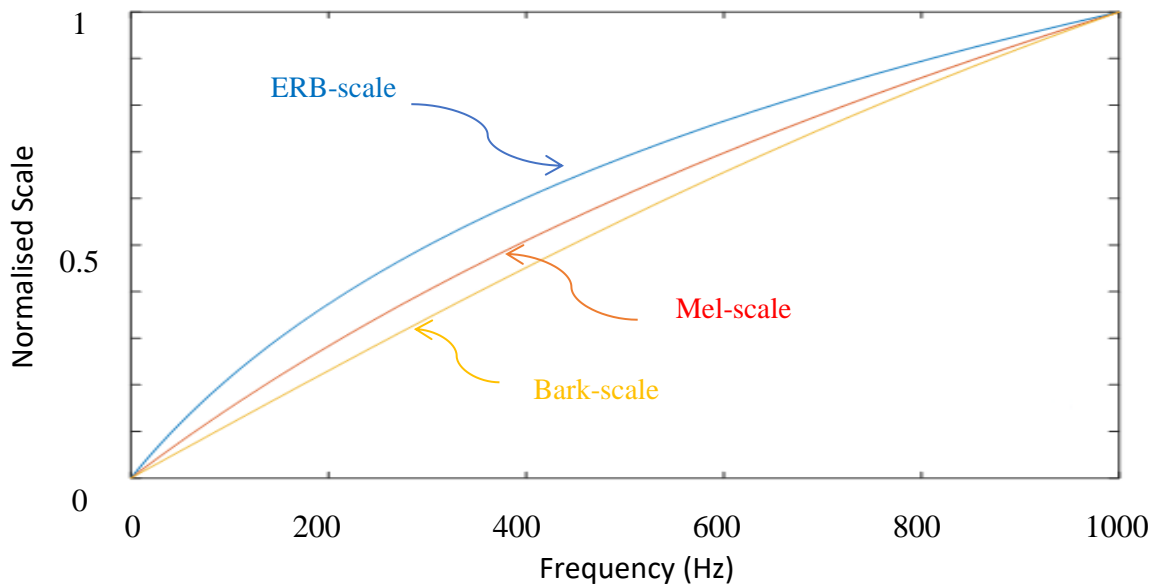


Figure 3.4 The normalized-scale curves of the Critical-Bands analysis. *ERB*, *Bark* and *Mel* nonlinear frequency domain scales on the range of the major part of the speech energy, i.e. under 1 kHz ([Giampiero Salvi website](#)).

From laboratories tests, the *ERB*-scale is expressed by the following approximated equation [134]:

$$ERB(f) = 24.7 \left(\frac{4.37f}{1000} + 1 \right) \quad (3-6)$$

where the *ERB*(*f*) is in Hz and “*f*” is the center frequency in Hz.

From laboratories tests, the *Bark*-scale is approximately expressed by the following equation [135]:

$$Bark(f) = 13 \arctan(0.00076f) + 3.5 \arctan \left(\left(\frac{f}{7500} \right)^2 \right) \quad (3-7)$$

where the *Bark*(*f*) is in Hz.

Also, from laboratory tests, the *Mel*-scale is expressed by the following equation [126]:

$$Mel(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (3-8)$$

where the *Mel*(*f*) is in Hz.

To enhance the PLP pre-process, *Hermansky*, *Morgan*, *Bayya* and *Kohn* in [136], then *Hermansky* and *Morgan* in [137] developed the **RelAtive-SpecTrAl** (RASTA) filter methodology. To describe the RASTA filter arrangement, inside the Log-Scale frequency domain, RASTA filter uses a specific bank of bandpass filters, then uses the removing of the Slow Channel Variations.

RASTA filter has pre-process on the cepstrum feature-based on the Log-Spectral and Cepstral-Domain filters. To detection of the utterance, the unvoiced and the silence periods of the input speech signal of RASTA filter should not be removed. Figure 3.5 is the simplified functional block diagram of the RASTA filter. To attain the best effectiveness from RASTA filter, it is located inside the main algorithms of PLP sequences, between the “Equal-Loudness Pre-Emphasis Filter” block and the “Nonlinearity Power Function” block; see Figure 3.6.

The input speech signal of this step is arranged in accordance with the previous step. Frame-by-frame, the overlapping-window frames input to the RASTA-PLP. The outputs are 13 corresponding coefficients (features) N_c , these represent each frame. The 13 extracted features have been chosen after the cross-checking of 8, 10, 13, 17, 24 and 32 features per frame.

For the frequency domain filtering, the *Mel*-scale, the *Bark*-scale and the *ERB*-scale have been checked; then the *Bark*-scale has been chosen because it provides the best output results for this chapter’s algorithm. In order to attain the flatness property of these filters, fourth-order *Butterworth*

filters were employed. The number of these filters are calculated by the referring to the formula that available in *Hermansky* literature [131, 136, 137]:

$$\text{Number of Filters} = [\text{ceil}(6 \text{ asinsh}(BW/600))] + 1 \quad (3-9)$$

where $\text{ceil}(\cdot)$ is the ceiling function that calculates the approximated upper integer, asinsh is the inverse hyperbolic sine function, and BW is the bandwidth of the input speech signal. For 16000 sample/s, number of these filters are 21; and 17 filters for 8000 samples/s. Since the speech signal is divided into N_f total overlapping-window frames, and each frame has its unique 13 features, the resulting features could be configured as $[13\text{-by-}N_f]$ array (the (b)/Figure 3.7).

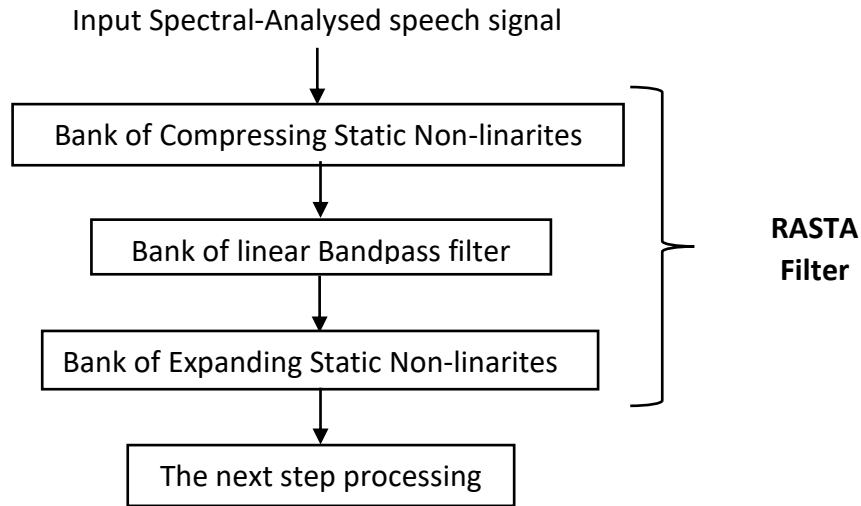


Figure 3.5 Flow chart of the simplified version of **RelAtive-SpecTrAl (RASTA)** filter [137].

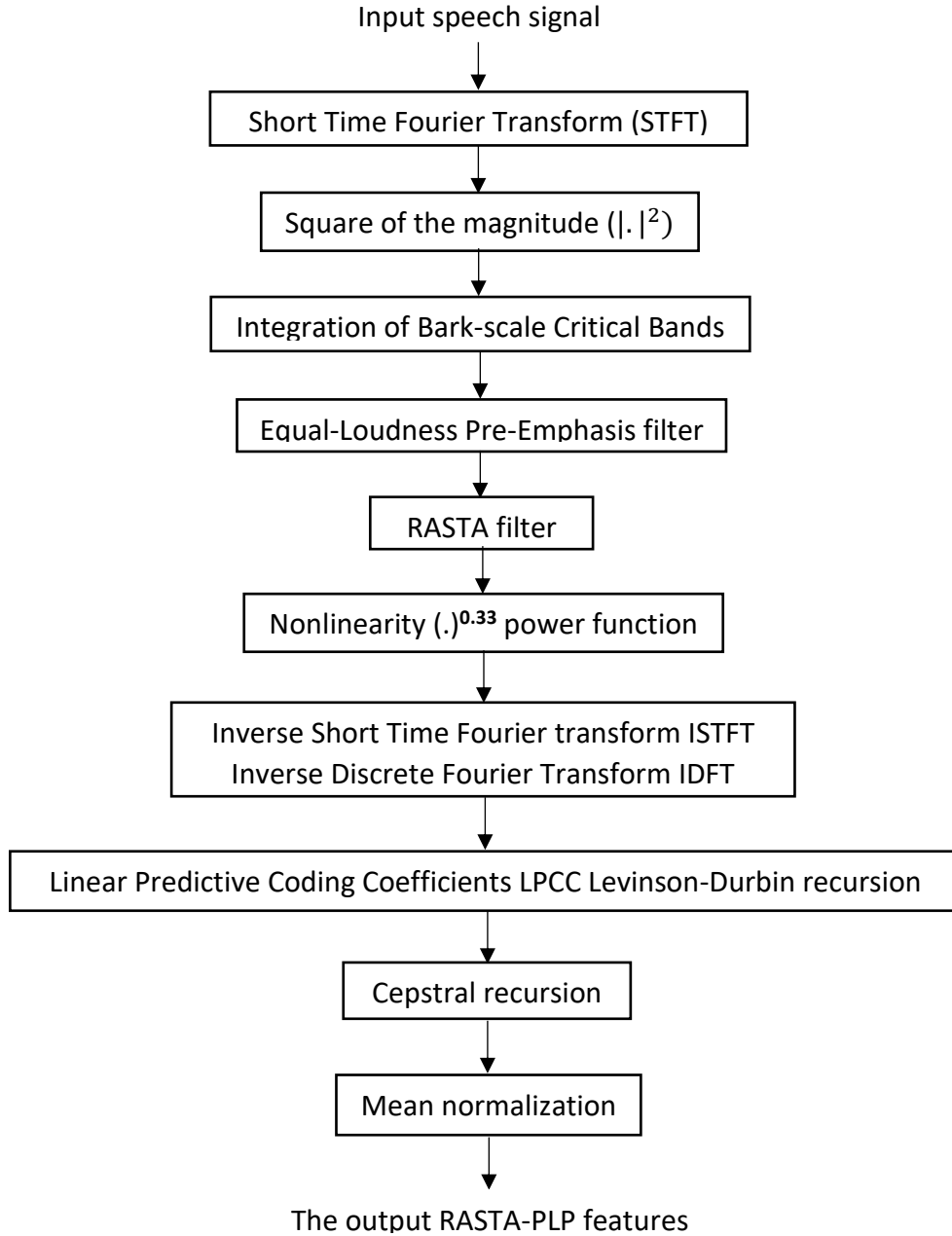


Figure 3.6 Flow chart of the RASTA-Perceptual Linear Predictive Coefficients (RASTA-PLP) to extract the audio features [136]. More details about the flow charts in Figure 3.3 and Figure 3.5 can be found on the literatures [131, 136, 137] of *Hermansky, Morgan, Bayya and Kohn*.

3.3.3 *k-means Clustering of the Features*

The next step of the algorithm is the clustering of the extracted features into 2 clusters and into 3 clusters. In the case of 2 clusters: the 1st cluster denotes to the dialogue speech features, the 2nd cluster denotes to the mixture speech features. The space of the resulting vector of the clustering

process should be the set $\{1, 2\}$. In the case of 3 clusters: the 1st cluster denotes to M speech features, the 2nd cluster denotes to FM speech features and the 3rd cluster denotes to F speech features. The space of the resulting vector of the clustering process should be the set $\{1, 2, 3\}$. The $\{1, 2\}$ and $\{1, 2, 3\}$ do not have any calculation values because they are the arbitrary labels of the clusters: {dialogue mixture}, and the clusters: {M, FM, M}.

There are many algorithms, techniques and algorithm which are used to cluster any data to specific (known or unknown) number of clusters. These algorithms are based on different approaches such as the connectivity-based clustering (the hierarchical clustering), the centroid-based clustering, the statistical distribution-based clustering and the statistical density-based clustering [138].

The centroid-based clustering algorithms are the k-means and the k-medoids. The k-means technique is an efficient and one of the simplest cluster analysis, where k is the known number of the required clusters. The k-means is a centroid-based approach which is performed after finding the k numbers of means (centroids) of these clusters, then sharing the input data according to the nearest distance to these centroids. **Appendix C** has been added to the thesis to expand the description.

The k-means has been described in the appendix A of the thesis, which includes a brief description of this well-known algorithm and the historical overview of the famous implementation algorithms (e.g. *Lloyd*-algorithm) [139]. In addition to these, the description presented the main problems which are posed by these algorithms and the recent solutions to overcome these problems [140, 141].

The output of k-means clustering is a vector of N_f elements (K -vector). The elements of the K -vector are either 1 or 2, for the clustering of the frames' features either the dialogue or the mixture speech (the (c)/Figure 3.7) The elements of the K -vector are either 1, 2 or 3, for the clustering of the frames' features either the M, the FM or the M (the (d)/Figure 3.7). Obviously, the figure shows that the major labels are false and the minor labels are correct. The subjective and the objective tests, of this step, denote that the above clustering are bad and the labels have a lot of errors. According to these tests, the above crude clustering does not have the enough capability for the proper detection. Instead of that direct crude use of that clustering, improvement(s) could upgrade this capability. The next paragraph and steps present the proposed improvements. The Figure 3.7, the Figure 3.9, Table 3.2 and Table 3.3 list the comparison for that initial clustering step with the next modified clustering.

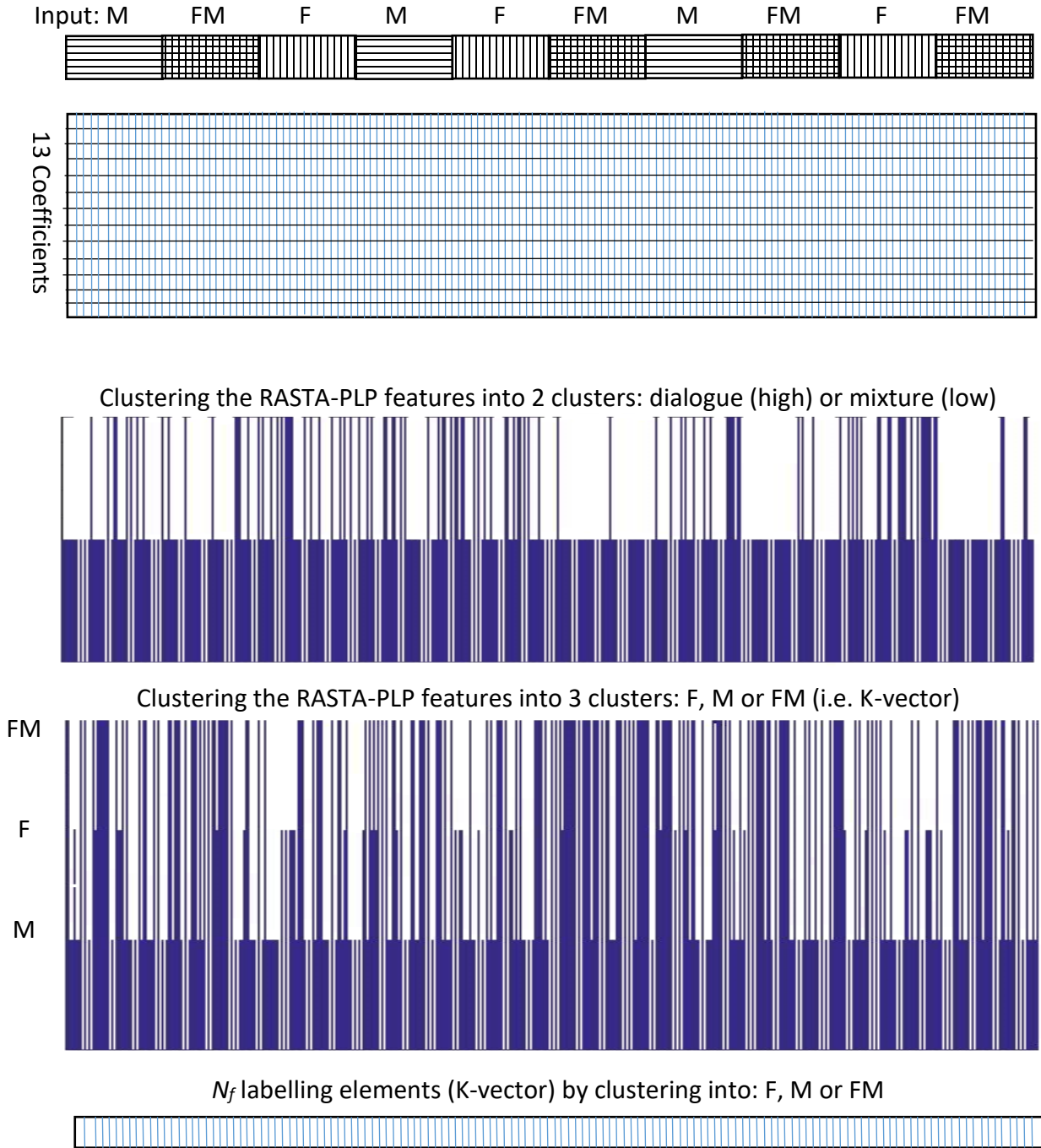


Figure 3.7 Audio features extraction, and Initial crude clustering. The (a) is the input spontaneous conversation. The (b) is an array which contains the $[13\text{-by-}N_f]$ features extracted by the RASTA-PLP. The (c) is the clustering of the features into 2 clusters: label-1 for the mixture and label-2 for the dialogue. The (d) is the clustering of the features into 3 clusters: label-1 for M, label-2 for F and label-3 for FM. The lower line is the K-vector which contains the k-means clustering results of the (d), i.e. into 3 clusters: M, FM or F. There are horizontal-axes time-domain relationships between all the sketches.

At first, the modification uses the clustering of the features to M, FM or M labels, i.e. the elements of the K -vector are 1, 2 or 3. Suppose they are the labels of M, FM or F respectively; see Figure 3.7 . Assume that the instances of the switching times, from any speech segment to its following speech segment, are known (this assumption is very important for the proposed modification). The period of each speech segment is prepared for 30 s/segment. The speech segments are 10, and each segment has 3000 frames, so the number of frames of F speech is 9000 (3×3000), and M has the same number of frames. FM has 12000 frames because there are 4 segments (4×3000). For each: M segment, FM segment and F segment, the Probability Density Function (PDF) are calculated. The calculation is done by counting of the chances of 1, 2 and 3 of their corresponding values in K -vector. This counting is the Histograms of the segment. The PDF of each segment is the per unit normalization of each histogram. Statistically, the PDFs are discrete and finite; Figure 3.8 illustrates the PDFs of the 1st, the 2nd and the 3rd segments of the conversation. It is easy to calculate the variances of these 10 PDFs. Obviously, there are a wide-range of differences between the variance of the mixture speech (FM) in comparing with the variances of the dialogue speech (F or M). The mixture has higher variance and the dialogue has lower variance. According to that wide range, the conclusion is the clustering of above features could be achieved successfully by the use of any well-known technique such as k-means or k-medoids. This clustering introduces the segregation task directly by using binary masks. The above conclusion has been investigated on a sample of 24 arbitrary female and male speakers. Since they are 24 speakers, and each conversation includes 2 speakers; the number of the investigated conversations are: $((24+1) \times 24/2) = 300$, (number of chances = $N \times (N+1) / 2$). A number of the successful conversations is 297 (99%) and the failed conversations are only 3 (1%). The (c) and the (d)/Figure 3.9 are the results of the clustering of the successful and the failed conversation respectively. The clustering is divided into 2 clusters according to the values of the variances: high or low. This conclusion is correct and excellent when the switching instants are known, but always these important instants are unknown and not easy to predict them. The first stage of the speaker diarization process is called the Speaker Segmentation. This stage can track-and-estimate the locations of these instants. The speaker diarization is only deals with input dialogue conversations. In this chapter case, the input is a hybrid conversation which contains both dialogue and mixture speech signals. Instead of the traditional speaker diarization, the above approach has been adopted for this research because it has the excellent ability for the segmentation of those spontaneous

conversations. The high efficiency performance is the main motivation for that. The tracking-and-capture of these instants is the key to the solution. The estimation of proximity of any actual instant is the suggestion which leads to very good results.

The differences between any estimated time and its actual time of occurrence, produces error. Accumulation of these errors reduces the overall efficiency of the system, but this reduction is acceptable according to the final tests.

To estimate those instants, the solution is by trying to find any instant in the range of 0.1 s to 3.2 s (10 to 320 frames). The lower limit is 0.1 s, because the resulting single error is negligible for a period of less than this limit. The upper limit is chosen 3.2 s by the trial-and-error. The second reason for this choice is the fact that if 2 or more switches during this period occur, the resulting error(s) are insignificant. This is because of the resulting errors are accepted in the estimation algorithm. Another reason for that is the fact that the duration period has been taken into account in the optimal formula that will be used in the next subtitle (the relationship is formulated as inversely proportional). Inside that range (0.1 s to 3.2 s), 32 switching times are suggested, then investigated carefully according to the machine-learning and the pattern-recognition basics and principles. Periodically, this algorithm is repeated to find the next switching-instant and so on to find the other switching-instants. The details are presented in the following optimization algorithm.

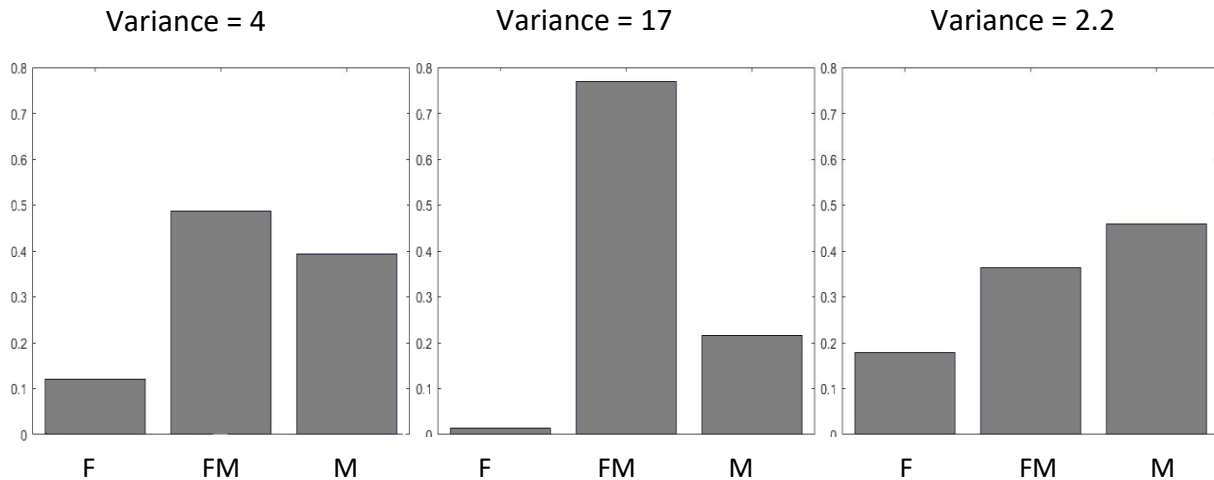


Figure 3.8 Specimens of three PDFs. Each one for the clustered labels of the extracted features of M (left), FM (middle) and F (right) speech.

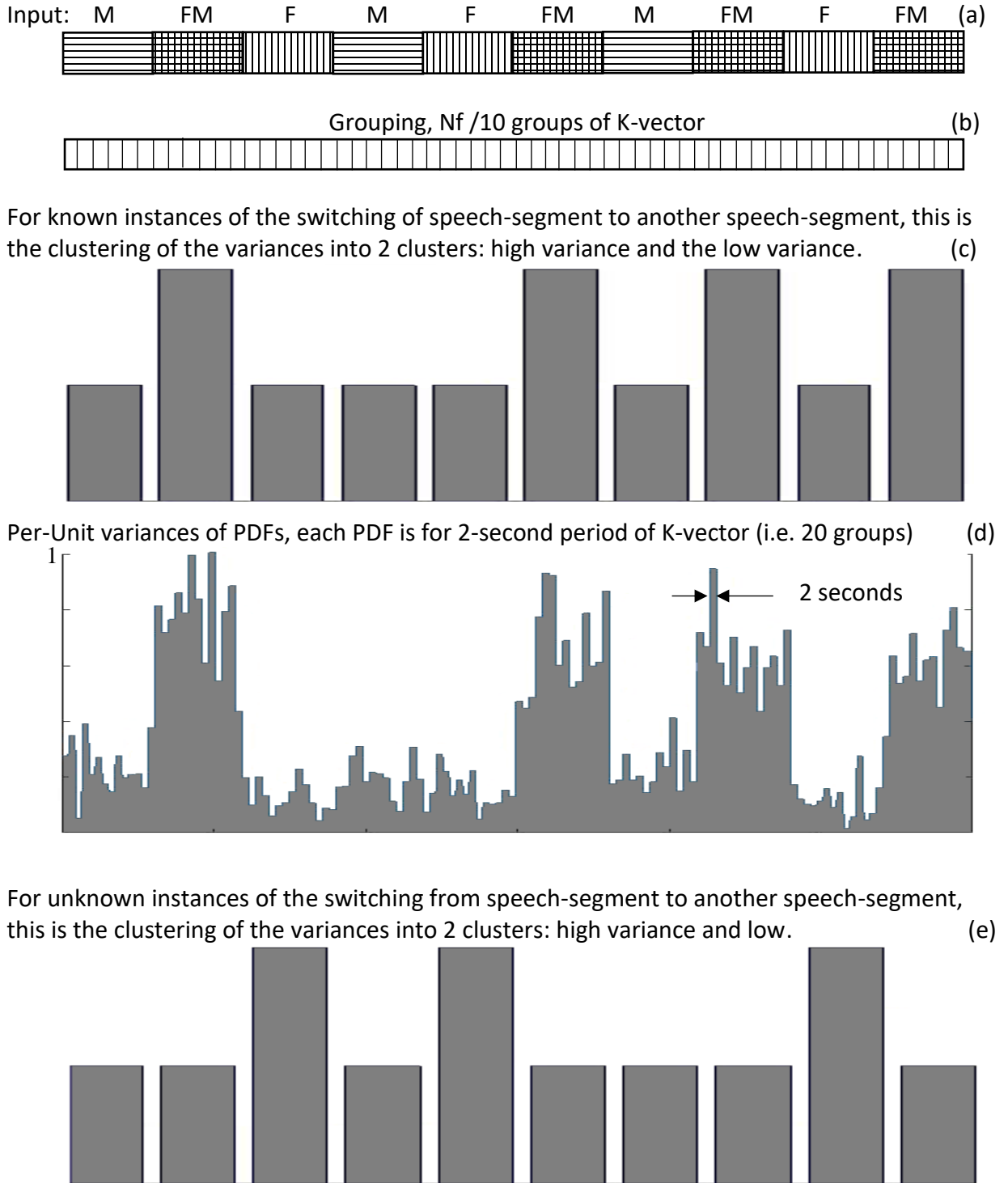


Figure 3.9 Grouping concept. Each 10 frames (0.1 seconds) is the fundamental group. The concept facilitates the mission of finding the switching instants. The (c) is the perfect clustering of the variances when the switching instants are known. The (d) for the supposed switching instants are regular each 2 second. The (e) is the worst clustering of the variances, when the switching instants are unknown. There are horizontal -axes time-domain relationships between all the sketches.

3.3.4 Groups and Statistical Variances

The following optimization algorithm, consists of several modification terms of the previous traditional clustering algorithm. These terms are sequential, but they are interlaced and interactive in the most cases. The following detailed description confirms these interlacing and the interaction. The output of the k-means algorithm is the K -vector of N_f elements in which each element is either 1, 2, or 3. In the case of switching instants in each 0.1 s period, the elements are collected in each 10 adjacent elements in one group (G_1), i.e. the group collects the labels of the 0.1-s period of speech. For each group, numerically, the PDF of these 10 values are calculated, followed by the Variance V_1 of the PDF. Group-by-group, this algorithm is done by the calculation of the PDF of each group then the variance of this PDF. Because number of the groups N_G is ($N_f/10$), number of the sequential variances N_V is ($N_f/10$).

These values of the variance are located on the first row of the array, which is called V-array (see the 1st row/Table 3.1 V_1^1 to $V_1^{N_f/10}$). Since this row contains the variances of periods of 0.1-s switching-instants, the row should be clustered into 2 clusters: the 1st for the dialogue and the 2nd for the mixture. The output of this clustering resides in the 1st row of D-array (it is the Decision-Array).

The weak point of the 0.1-s period, is the fact that the number of the statistical trials is 10, which is not enough to cover the statistical requirements for the proper accuracy. To increase the accuracy, and to try another location for the switching instant, the previous algorithm is repeated with 20 adjacent elements of each group; i.e. by the dividing of the K -vector into ($N_f/20$) groups (G_2). For each G_2 , its PDF is found then its variance is calculated. The number of the resulting variances for all the V_2 groups are ($N_f/20$). These variances V_2 values should be located in the second row of the V-array, but this number is a half of the column number of the V-array. Both the previous ($N_f/10$) variances and these ($N_f/20$) variances cover the duration of the total speech period. Thus, the correct configuration of the second row of V-array is achieved by repeating of each variance two times, and locating them, one adjacent to the another (see the 2nd row/Table 3.1 V_2^1 to $V_2^{N_f/20}$). Since this row contains the variances of periods of 0.2-s switching-instants, the row should be clustered into 2 clusters: the 1st for the dialogue and the 2nd for the mixture. The output of this clustering resides in the 2nd row of the D-array.

In order to increase the accuracy of the above algorithm and try other switching instants, a number of the elements of each G_3 group is 30. By the dividing of the elements of the K -vector to ($N_f/30$),

G_3 groups are produced. The PDFs are calculated for the groups one-by-one, then the variances V_3 of the groups one-by-one are calculated. The $(N_f/30)$ variances should be located in the 3rd row of the V -array. Because number of them are 1/3 of the first-row number, the variances are corrected by the repeating of each variance three times and locating the repeated values one adjacent to the others (see the 3rd row/Table 3.1 V_3^1 to $V_3^{N_f/30}$). Since this row contains the variances of periods of 0.3-s switching-instants, the row should be clustered into 2 clusters: the 1st for the dialogue and the 2nd for the mixture. The output of this clustering resides in the 3rd row of the D -array.

The same algorithm is done for the G_4 , the G_5 , the G_6 , ... etc. They have 40, 50, 60, ... etc. variances (V_4 , V_5 , V_6 , ... etc.). Number of these groups are $(N_f/40)$, $(N_f/50)$, $(N_f/60)$, ... etc. They are repeated 4, 5, 6 and subsequent times. They are located in the 4th and subsequent rows of the V -array, see Table 3.1. The repeated values are located one in the adjacent of the others (the 4th row, the 5th row, the 6th row and so on /0: V_4^1 to $V_4^{N_f/40}$, V_5^1 to $V_5^{N_f/50}$, V_6^1 to $V_6^{N_f/60}$, and so on), see the Table 3.1. Since these rows contain the variances of periods of 0.4-s, 0.5-s, 0.6-s, and so on of the switching-instants, each row should be clustered into 2 clusters: the 1st for the dialogue and the 2nd for the mixture. The outputs of these clusters reside in the 4th row, the 5th row, the 6th row and subsequent rows of the D -array.

To try other locations of switching-instants, and to increase the statistical accuracy, the previous algorithm should be continued until the adequate upper limit is reached. According to the nature of the speech, 3.2 s is chosen after the trial-and-error. Another reason for that limit is the fact that the statistical accuracy tends to reach its saturation state near 320 trials. Another reason for choosing this number is the fact that the spoken word does not consume more than several seconds, approximately [142], so the errors of the clustering (if occurring) should not spread to the forward(s), and/or the backward(s) nearest words which may be clustered correctly. In case of less than 0.1-s or more than 3.2-s speaking period, the Hierarchical Clustering Scenarios could redress such case by the splitting or the merging of the group(s). For all these reasons, the smallest group has 10 elements of the K -vector and the largest group has 320 elements of the K -vector. For the largest-group set, number of groups is $(N_f/320)$. Number of variances V_{32} is $(N_f/320)$, and they are repeated 32 times to reside in the row number 32 of the V -array (see the 32nd row/Table 3.1 V_{32}^1 to $V_{32}^{N_f/320}$) [70-72].

Table 3.1 V-array arrangement. The Decision-array (which has the same dimensions) by the clustering of each row of V-array into 2 clusters: High variances and Low. This algorithm is done for the 32 orders of groups (32 rows).

V_1^1	V_1^2	V_1^3	V_1^4	V_1^5	V_1^6	V_1^7	V_1^8	V_1^9	V_1^{10}	V_1^{11}	V_1^{12}	$V_1^{N_f/10}$
V_2^1	V_2^1	V_2^2	V_2^2	V_2^3	V_2^3	V_2^4	V_2^4	V_2^5	V_2^5	V_2^6	V_2^6	$V_2^{N_f/20}$
V_3^1	V_3^1	V_3^1	V_3^2	V_3^2	V_3^2	V_3^3	V_3^3	V_3^3	V_3^4	V_3^4	V_3^4	$V_3^{N_f/30}$
V_4^1	V_4^1	V_4^1	V_4^1	V_4^2	V_4^2	V_4^2	V_4^2	V_4^3	V_4^3	V_4^3	V_4^3	$V_4^{N_f/40}$
....
....
....
V_{32}^1	V_{32}^1	V_{32}^1	V_{32}^1	$V_{32}^{N_f/320}$

Since this row contains the variances of periods of 3.2 s switching-instants, the row should be clustered into 2 clusters: the 1st for the dialogue and the 2nd for the mixture. The output of this clustering resides in the 32nd row of the Decision-array. The results of the above arrangement are the [32-by-($N_f/10$)] V-array, as shown in the Table 3.1 and the [32-by-($N_f/10$)] Decision-array. The V-array contains the variances of the groups and Decision-array contains their clustering labels: 1 or 2.

3.3.5 Optimizing the Groups

In most cases, pattern recognition and machine learning methodologies are facing the following two main challenges: The first challenge is the choice of the best algorithm which extracts the proper features, which are achievable of a specific research. In this research chapter, this challenge has been overcome by trial-and-error. Technique-by-technique, the results of these techniques are tested subjectively then objectively. After the checking and the testing of several techniques, successfully, the RASTA-PLPC satisfies the conditions of this research, so it has been chosen.

After the choice of the specific technique, the second challenge that the researchers are facing is the fact that, the performances of the features fluctuate and are not similar. Certain features have good performance, but the other parts do not have, relatively. This problem becomes very complicated when the good part for specific frames is bad against other frames, and vice-versa. The goodness and the badness levels are deterministic, and relatively evaluated. The decisions on

the adoption of the good part and the omission of the bad parts are not realizable many times because this process is dynamic. Sometimes, the static process of the speech is better than the dynamic due to the quasi-stationary stochastic behavior of the speech. Instead of choosing the best features and omitting the worst features, for this research chapter, choosing the best groups and omitting the worst groups, is the used algorithm. The next paragraph describes the way to the optimization of the features, which are supported by the main PR & ML properties [70-72].

Suppose there are two patterns and each one is represented by its PDF. The first has n data with their n probabilities, and the second has m data with their m probabilities. Each PDF has its mean (centroid), i.e. AV_1 is the average value of the first and AV_2 for the second. Each PDF has its variance value too. Graphically, part of the data of each pattern is located on the Left-Hand-Side and the other part is located on the Right-Hand-Side of its mean. The Distance between these two centroids (D_{cent}) determines the common overlapping area between the data of these patterns. Basically, such overlapping area is inversely proportional with the D_{cent} , i.e. the largest distance is the less overlapping area and the nearest distance is the largest overlapping area; Figure 3.10 illustrates this relationship graphically.

According to the basics and the principles of PR & ML, the overlapping common area between the two patterns is the main factor which evaluates the *Goodness* of the recognition-decision between their patterns [2, 19]. Obviously, the right decision for the clustering analysis depends strongly with this overlapping area. Roughly, the clustering goodness has a direct relationship with the distance between the centroids of the processed data:

$$Goodness \propto D_{cent} \quad (3-10)$$

According to the statistical principles, the variance of a PDF pattern represents the spread of that pattern. Basically, the overlapping area is directly proportional with the variances of the PDFs of these patterns, i.e. the largest variances are the largest overlapping area, and vice-versa; Figure 3.11 illustrates this relationship graphically. Roughly, the clustering goodness is inversely related with the variances of the PDF distribution of these data [70-72]:

$$Goodness \propto 1/V \quad (3-11)$$

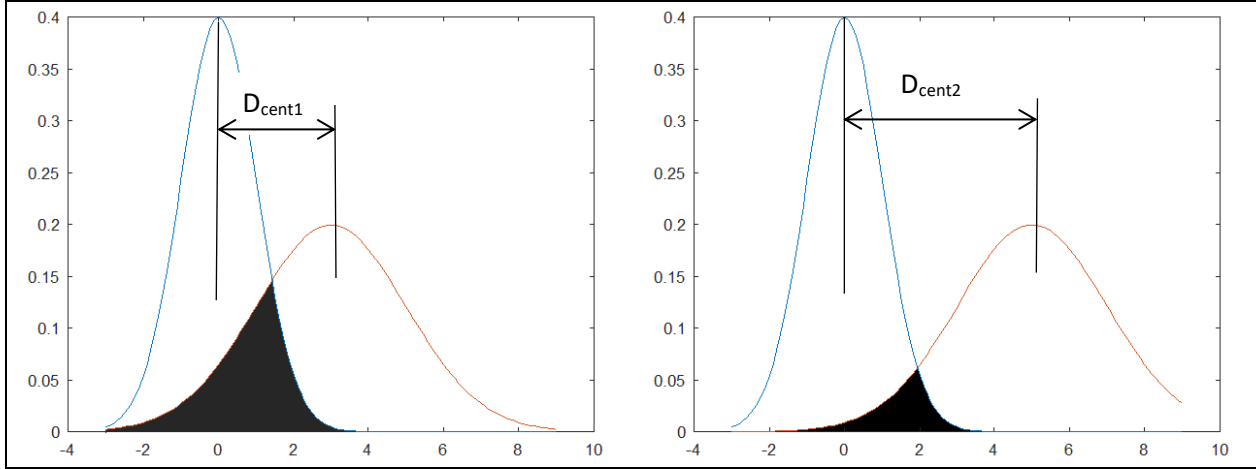


Figure 3.10 Sketches illustrate the effect of the distances of the centroids of two PDFs. The patterns are same, but the difference is the distances between their centroid. The large distance is the less common overlapping (black area) & vice-versa.

In addition to the effects of the distances between the centroids and the variances of the patterns PDFs, the group period τ has clear effect. At first, the group-period of less than 0.1 s (10 frames) is not reasonable as a switching time and not acceptable statistically because it has very small number of trials. The increase of this period provides an enhance statistical indication (more of the trials), but for more than 3.2 s, the enhancement tends to reach its saturation asymptotic. On the other hand, the wide time of group-period τ has an inverse effect on the recognition-decision because the continuously spoken words do not consume more than this range of time.

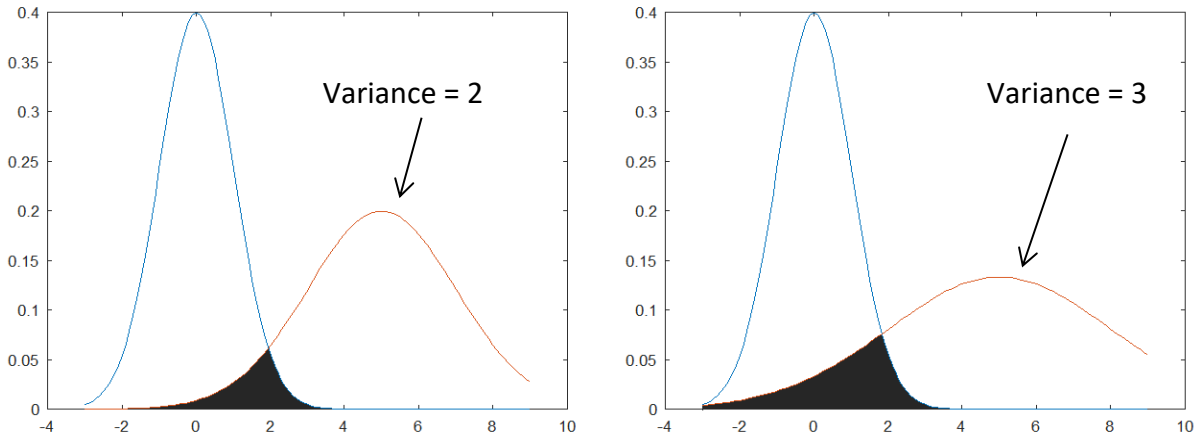


Figure 3.11 Sketches illustrate the recognition between two distribution patterns. The difference is the variance of one of them. The high variance is the large overlapping (black area) & vice-versa.

Another reason is the fact that the chances of occurrences for switching periods of less than 3.2 s exist. Experimentally [70-72]:

$$Goodness \propto 1/\tau^\beta \quad (3-12)$$

where β is between 0.3 to 0.5 values. According to the effects of the D_{cent} , the V and, the τ , the equations (3-10), (3-11) and (3-12) can be summarized to produce the overall relative relationship of the goodness of the recognition-decision versus all these three parameters:

$$Goodness \propto \frac{D_{cent}}{V \times \tau^\beta} \quad (3-13)$$

Because the k-means clustering is depending entirely on the distances from the data with the centroids of the clusters, the locations of these centroids are the main step for its manipulating. The D_{cent} is the Cartesian distance between these centroids. The 32 corresponding variances could be calculated from the set configured inside the V-array.

For the above 32 choices of grouping, the corresponding calculations of equation (3-13) are the relative goodness for the recognition between these two patterns versus the duration of these groups. These results are changing from speaker to other speakers. As well as this, these results are changing for the same speaker if the speaker is speaking different paragraphs. Thus, the above optimization is a dynamic process. The decision, of which groups are good and which groups are bad, changes from speaker to speaker and from speech to speech. Using $\beta = 0.4$, the formula (3-13) has been applied on the conversation between F and M of TIMIT standard speech library (see Figure 3.1) [20]. The relative goodness of the above algorithm is per-unit calculated. The results of these calculations are shown in Figure 3.12.

To utilize the useful good groups and avoid the harmful bad groups, these groups should be split into two types: good and bad. This algorithm has several advantages. The first advantage is the estimation of the switching time. The second is the avoiding, indirectly, of the bad group of features by neglecting the groups which contain the largest number of the harmful features.

Unfortunately, there is no specific mathematic derivation to specify the goodness level of such different level of goodness, so the best for that is the threshold that borders them. The k-mean algorithm has the simple and efficient ability to draw this clear borderline, i.e. by the clustering of the formula (3-13) results into 2 clusters: one contains the good groups and the other contains the bad groups. This is attainable by the calculations of the 32 choices of groups on the formula (3-13),

then the drawing of that borderline. The resulting good (Rich) useful groups are the 19 groups: {all the groups: from G3 to G20, and G22}. The resulting bad (Poor) harmful groups are the 13 groups: {G1, G2, G21, and all the groups: from G23 to G32}. The good groups are useful to re-process the above algorithm again efficiently. The bad groups are omitted, because they are harmful to that re-process [70-72]. Figure 3.12 shows the line which is the border between them.

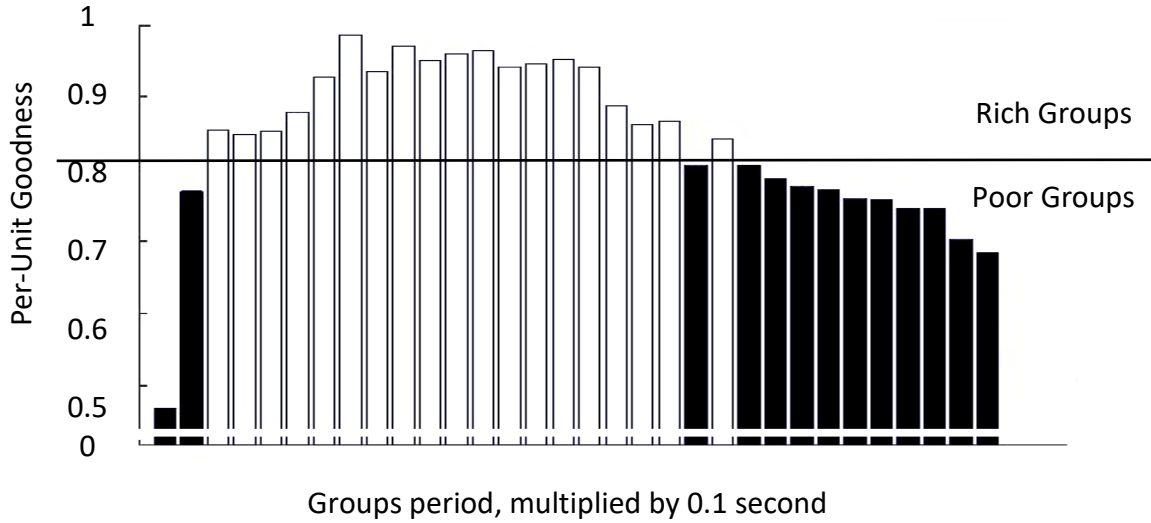


Figure 3.12 Per-Unit Goodness of the recognition between the patterns of the variances of the audio features of spontaneous conversation between F&M. The horizontal line is the border the useful rich (white) groups and the harmful poor (black) groups.

3.3.6 Re-clustering

The grouping step produced two arrays corresponding to the original features of the conversation. The 1st array, which is the V-array has been used for the previous “Optimization of the groups” step. This step is an optimization process which involves discarding the harmful groups and retaining the useful. The second array is the Decision-array which has the discrete values of the clustering of each row alone. The clustering is into 2 clusters: the 1st cluster is labelling the dialogue speech groups, which are the speech of the female F or the male M. It is not possible to conduct speaker diarization of this dialogue speech. The 2nd cluster is labelling the mixture speech groups, which are the speech of the female and the male simultaneously FM. Similarly, speech separation is not possible for this mixture speech. The Decision-array has $[32\text{-by-}(N_f/10)]$ dimensions. Each cell of this array is a decision of 0.1-s period of the conversation, either the dialogue speech or the mixture speech. For each 0.1-s group, there are 32 decisions. Sometimes, the following

reconciliation between these 32 decisions occur: {clear majority is the dialogue and clear minority is the mixture, clear majority is the mixture and clear minority is the dialogue}. Other times, the following conflict between these 32 decisions occur: {majority is the dialogue which approximately equals to minority of the mixture, majority is the mixture which approximately equals to minority of the dialogue}. By the averaging, and then the comparing with the midpoint between them (the threshold), the simple voting process is used to assign each group as a dialogue speech or a mixture speech [70-72].

The result is an acceptable segregation process. However, in a stream of correct decisions, there are fragments of false decisions. In order to minimize these false decisions as little as possible, the standard hierarchical clustering scenarios could be used.

3.3.7 Hierarchical Clustering Scenarios

There are two general hierarchical clustering scenarios, the Top-Down (Divisive) and the Bottom-Up (Agglomerative) scenario. Depending on its application, each scenario has its specific technique. The technique achieves the refurbishing job of the few unidentified segments among the most identified neighborhood, or vice-versa [46, 143-145].

To explain the job of these scenarios and the knowhow of performing that task, suppose there are known number of speech segments (N_{ts}). The segments are the speech of different known or unknown number of speakers. The scenarios are auxiliary techniques which aid the main identification algorithm for assigning of these segments to their corresponding speakers. In case of the top-down (divisive) scenario, the start point is by the choice of more than one segment (N_{ss}) and supposing that all these segments belong to a specific speaker. Segment-by-segment, top-down scenario tries to eliminate the undesired segment from the chosen segments, and keep the desired segment with the chosen segments. That scenario should apply to another set of segments to eliminate the undesired and keep the desired. Set-by-set of segments, the scenario reduces the number of the unidentified segments to number of new segments N_{es} (less than the starting number). The algorithm could be repeated many times till the reaching of a boundary condition criteria (e.g. the Bayesian Information Criterion BIC [146] and the Viterbi-Stopping [147]). For the top-down (divisive) scenario, the number of the starting segments are $N_{es} \leq N_{ss} \leq N_{ts}$. The lower side in the Figure 3.13 illustrates the typical scheme of the top-down (divisive) scenario.

In case of the bottom-up (agglomerative) scenario, the start point is by choosing N_{ss} of less than

the total number the segments N_{ts} . Supposing that all these N_{ss} segments belong to a specific speaker. Segment-by-segment, bottom-up scenario tries to merge the desired segment to the chosen segments, and keeping the undesired segment outside the chosen segments. That scenario should apply to another set of segments to merge the desired and reject the undesired. Set-by-set of segments, the scenario reduces the number of the unidentified segments to a number of new segments less than the starting number. The algorithm could be repeated many times till the boundary criteria of such processing is reached. For the bottom-up (agglomerative) scenario, the number of the ending segments are $N_{ss} \leq N_{es} \leq N_{ts}$.

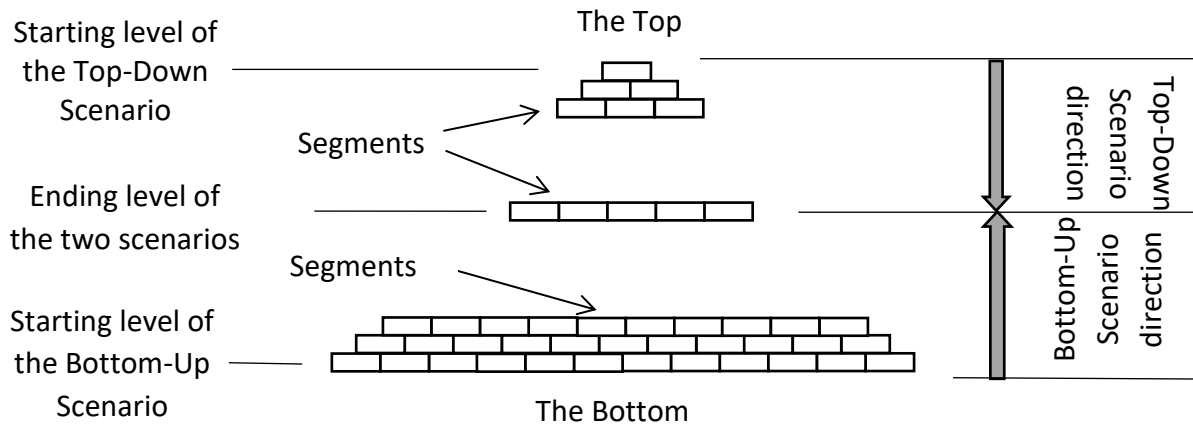


Figure 3.13 Typical scheme of the two Hierarchical Clustering Scenarios: Top-Down (Divisive) and the Bottom-Up (Agglomerative).

The upper side in the Figure 3.13 illustrates the typical scheme of the bottom-up (agglomerative) scenario [46, 143-145]. For the speaker diarization and the overlapped-speech detection, the bottom-up scenario is used more than the top-down scenario, because the bottom-up has more achievability [46].

Inside the set of segments, there are time-domain neighborhood between most of these segments. Because the segments are sequentially in the time domain, they are neighbors. Those neighborhood merits are between the eliminated/the merged segment and the other segment(s) of the chosen set. For the eliminating/the merging of the bottom-up/top-down scenarios, this merit is very significant for increasing the accuracy of the eliminating/the merging process. In addition to the accuracy, the neighborhood merits reduce the required time of processing and reduce the complexity of that process. Sometime the set of the chosen segments are compound, i.e. some of them are adjacent

neighbors but the others are not.

In this research segregation process, instead of the segments, the groups should be processed. Thus, the groups are identified either as the dialogue speech signal or the mixture speech signal.

The scenarios are the auxiliary techniques which aid the main clustering algorithm. By comparing between the process with and without the scenarios aid, there is an improvement in the final testing results. These improvements are different from a specific speaker to another speaker and from specific segment of speech to another segment. At the best scenarios efficiency, the improvements are not exceeding the 10% of the total achievement. The false decisions of the previous clustering are minors and reside inside the major correct decisions. They look like the fragments in-between the correct sequences. For these false fragments of groups, the neighborhood merit, almost, correct them perfectly. The difficult false groups are the groups near the switching instants. The Left-Hand-Side neighbor speaker of such instant is not the Right-Hand-Side neighbor speaker.

The Figure 3.14 is the flowchart of this chapter algorithm.

3.4 Experiments

To investigate their actual abilities, the algorithm has been tested step-by-step, carefully taking account of most possibilities. The number of the speakers taking part in the required conversations are 24 speakers. 10 of the speakers are females and 14 speakers are males. The source of 2 speakers is the TIMIT standard library for the speech and the audio DSP [20]. One of the TIMIT speakers is a female and the other is male (the F and the M, which are used in the description text).

The rest of the speakers are narrators; who are, arbitrarily, chosen from the well-known websites related to the audio books. The narrators of these books are famous (e.g. Dick Estell), and have deterministic and well-defined speech. For each one of these speech databases, the long periods of silence are removed, but the short periods of silence are not removed because the RASTA-PLP requires such short silences. Long period silence could cause the following conflict against the algorithm. When the two speakers are talking simultaneously (mixture speech), if one of them is talking continuously, but the is taking some time and remaining silent for a long period, the fashion of this mixture speech is mixture sometimes but dialogue other times. The period of the silence should be less than the longest neglected group.

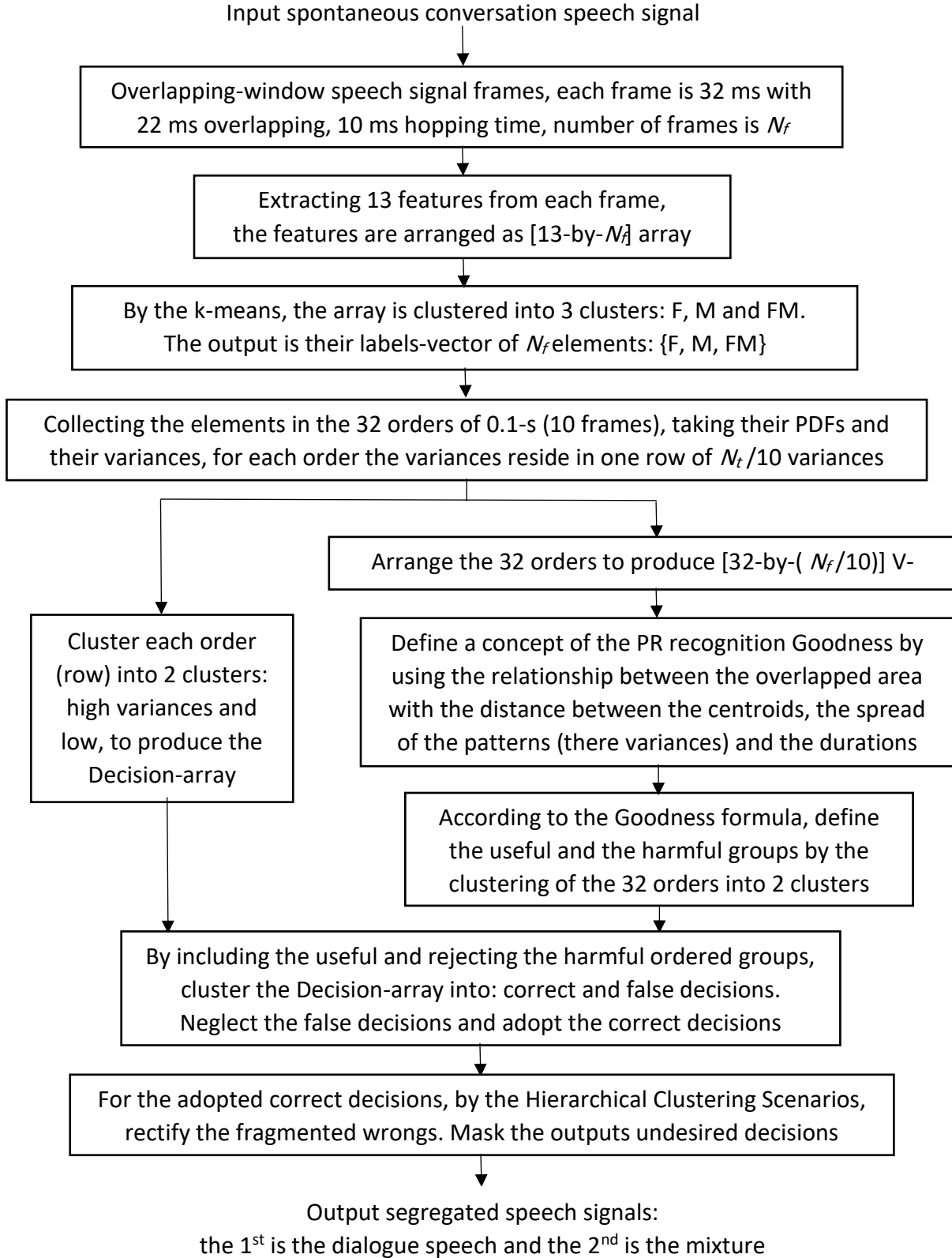


Figure 3.14 Flowchart of Chapter 3 algorithm.

Because the format of the standard PCM is without any compression or expanding of the speech, the database is formatted as (*FileName.wav*) of the PCM. For the cross-checking, the sampling rates of 16000 samples/s and 8000 samples/s are used. The resolution of each file is 32 bit/sample, but it has been reduced to the accepted resolution of 16 bit/sample.

The conversations have the same sequence, which is: F, FM, M, F, M, FM, F, FM, M then FM. The F is the speech of the female alone (the dialogue speech), the M is the speech of the male alone (the dialogue speech) and the FM is the speech of the female and the male simultaneously (the mixture speech). The conversations are prepared using the above sequence of speech segments (F, M or FM). The period of each segment is 30 s, so the conversation is 10 times 30 s, i.e. 300 s (five minutes).

In order to avoid any power-normalization problem during the processing, the segments have been power-normalised by the comparing of these segments with the standard energy of the segments of the F speech recording (the female of the TIMIT [20]).

For the audio files, the Audacity environment has been used continuously for the preparation of these audio files, for the quick and the subjective tests during the running of the executable files and for the final testing and playing of the resulting speech files. For the ASCII-code text, the notepad++ has been used for the editing of the source (*FileName.m*) files and the other data editing. For the ASCII-code tabling, MS-Excel has been used for the arrangement of the required database tables. The main speech-DSP is implemented by the using of the MATLAB environment, which is the most powerful DSP and statistics environment.

After the preparation of the required files, the implementation of the algorithm starts by the writing of the *SourceFile.m* for the using of the MATLAB DSP environment. In order to execute the first step of the algorithm, the required *FileName.wav* speech files should be read from its physical location, i.e. the conversation speech is fetched. The conversation speech signal is divided by the windowing-framed with 32 ms. The hopping of the frame is 10 ms. These frames are processed by the RASTA-PLP technique to produce 13 features each frame. The total number of frames are N_f for the conversation. The output of the RASTA-PLP implementation is the feature array with [13-by- N_f] elements; the (b)/Figure 3.7 Initially, these features have been investigated by clustering into 2 clusters: the dialogue speech and the mixture speech (by the k-means clustering algorithm). The out is a vector of N_f elements, each element is either 1 or 2. For the cross-checking, that investigation is repeated by clustering the features-array into 3 clusters: The F speech, the FM

speech and the M speech. The output is the K-vector of N_f elements, each element is either 1, 2 or 3; the (e)/Figure 3.7. The results of these two vectors have been plotted in the (c) and the (d)/Figure 3.7. Sequentially, the above algorithm is done for all the 300 FileName.wav files.

3.5 Result and Test

Arbitrary, 10 female and 14 male speakers are chosen. Since the female speakers are 10 and the male speakers are 14, the number of the prepared conversation files are 55 files for the F with F speakers $((n+1) \times (n/2))$, 105 for the M with M speakers, and 140 for the F with M speakers, i.e. the total is $55+105+140=300$ conversation files. To have the same evaluation balance, all the 300 conversation files have the same sequence of speaking. Inside each file, the chances for the first speaker is the same chance of the second (the above details confirm the similarity chances for both speakers). The resulting outputs are 300 pairs of speech, 300 for the dialogue and 300 for the mixture. According to subjective tests of these outputs, at a glance, the results are excellent. The errors of dialogue speech inside the mixture speech period and vice versa are very limited. The locations of these errors are only on the switching instants at the LHS and RHS around these instants. Most of these errors are insensitive. Occasionally, some errors are sensitive when the testers hear less than 0.5 s of errors around these instants. The upper waveform of Figure 3.15 shows the input conversation of the F and M of TIMIT. The middle waveform is output dialogue speech and the lower waveform is the mixture [20].

All the 300 conversations are checked subjectively. The responses of the testers (listeners) are highly positive. When the listeners are asked about the heard errors, they answered that the errors are only audible at some switching instants, and the worst-case error is for fraction of second. They said that the errors are around the switching instances only [70-72]. Output signal waveforms, subjectively denote non-existent errors outside the switching instances. And there is a lack of errors around the switching instances.

For each conversation, to evaluate the steps of the algorithm, objective tests are done for the 4 main steps of the algorithm. The tested 4-steps are:

1. The clustering of the features into 3 clusters: F, M and FM.
2. The clustering of the features into 2 clusters: dialogue (F or M) and mixture (FM).
3. The clustering of variances into 2 clusters: high and low.
4. The overall algorithm test.

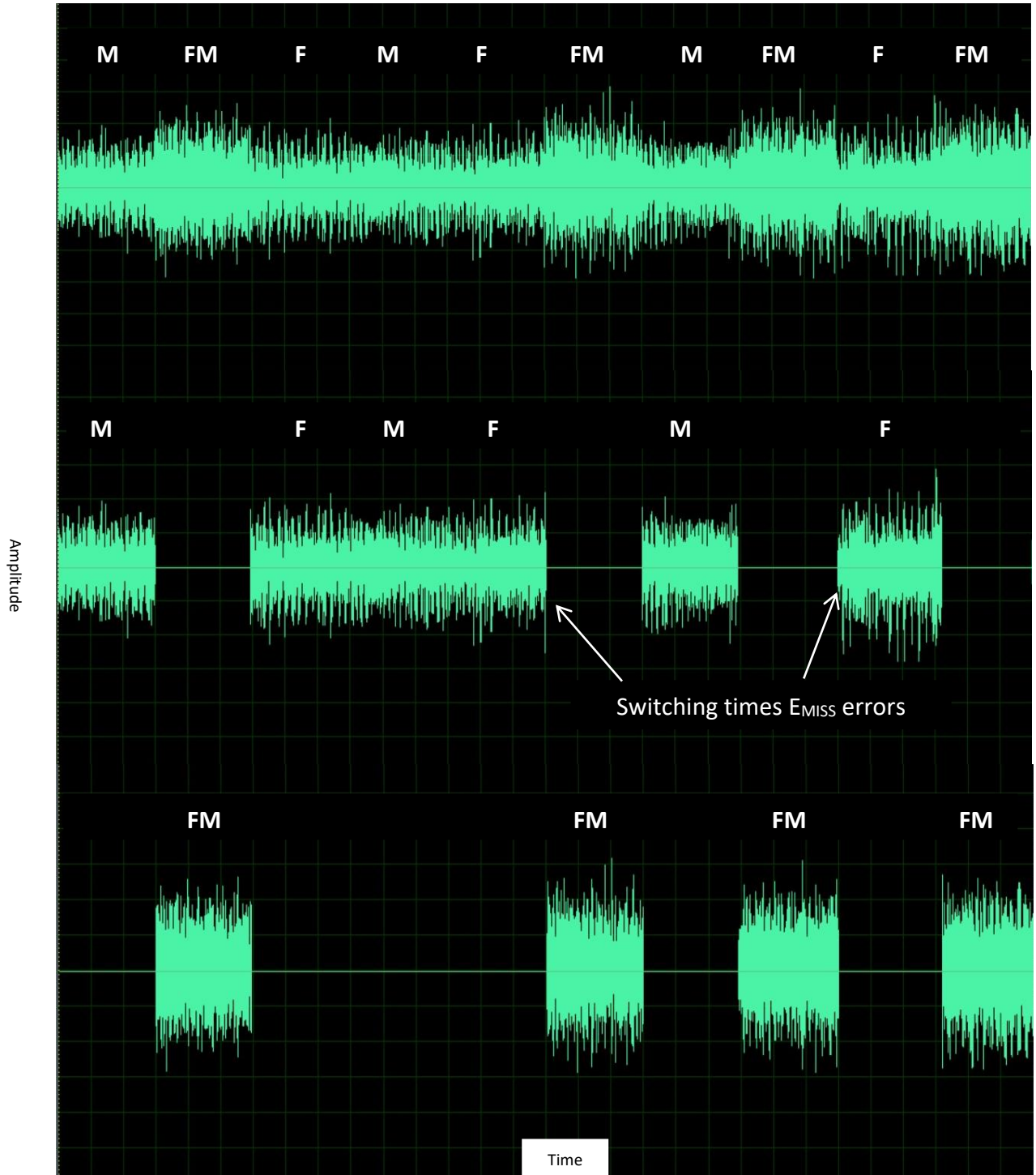


Figure 3.15 The waveforms of the implementation of Chapter 3 algorithm. The upper waveform is the input spontaneous conversation between the Female F and the Male from TIMIT [20] standard audio & speech library. The middle waveform is the output dialogue speech of the algorithm; its E_{MISS} is 0.5%. The lower waveform is the output mixture speech of the algorithm; its E_{FA} is 2.2%. The E_{OVL} is 1.2%. There are horizontal-axes time-domain relationships between all the sketches.

The Diarization Error Rate *DER* is the objective test that is calculated for these steps. Against each conversation, the *DER* of the each above step is calculated (**Chapter 1/1.10 Subjective Test versus Objective Test**).

The results of the *DER* [27, 122] are per unit, but to increase clearance of the evaluations, percentage rates (%) are used, so the *DER* of all the next calculations are multiplied by 100%.

The experimented and tested spontaneous conversations are categorized into:

- **Female with Female (FF).**
- **Female with Male (FM).**
- **Male with Male (MM).**
- The **All** conversations (**All**).

For each category, the resulting *DERs* are different, so the average value of these *DERs* is the final indication of each category during the specific step. In case of the All conversations, the following calculations are done:

1. The Average *DER* of FF category is multiplied by 55, the Average *DER* of FM category is multiplied by 140, and the Average *DER* of MM category is multiplied by 105. The 55, the 140 and the 105 are numbers of the FF, the FM and the MM conversation, respectively.
2. Sum the results of the above 3 multiplications in 1.
3. Divide the sum by 300 (number of the total conversations =55+140+105=300).
4. The resulting value is the average *DER* of the All conversations.



For each one of FF, FM, MM and All category, there are three overlapped-speech detection errors:

- **E_{MISS}**: When the detection suggests a mixture speech as a dialogue speech. This error is called Missed-Speech Error.
- **E_{FA}**: When the detection suggests a dialogue speech as a mixture speech. This error is called False Alarm Rate.
- **E_{OV}**: The overall error of the overlapped-speech detection. This error is called Overlap Speaker Error.

For this chapter experiments, during each conversation the total F periods and the total M periods are equal but the total FM periods are not. Since each conversation consists of 10 segments and each segment has the same duration of any other segment (30s/segment), number of the segments is used instead of the period's time in the calculations. The mixture speech has 4 segments (FM segments), but the dialogue speech has 6 segments (3 segments for F plus 3 segments for M). To

calculate E_{OVL} rightly:

1. The average E_{FA} of each category is multiplied by 4 which is the number of mixture segments and the average E_{MISS} of each category is multiplied by 6 which is the number of dialogue segments.
2. Sum the above 2 multiplication results of 1.
3. Divide the sum by 10, which is the total number of the segment per each conversation.

According to the above calculations for the *DER* objective test, for each spontaneous conversation, there are two resulting error rates. Instead of the tabulating of the all 300 calculated errors, their averages are remarked in the Table 3.2 and the Table 3.3. For the conversations: FF, MM and FM, these errors denote that the crude clustering has a lot of false decisions and the segregation efficiency of them are low. For the clustering into 3 clusters, the average error of 300 speakers of the mixture speech inside the dialogue speech and the dialogue speech inside the mixture speech is **34.7%**. For the clustering into 2 clusters, the average error of 300 speakers of the mixture speech inside the dialogue speech and the dialogue speech inside the mixture speech is **22.5%**, more details are listed in the Table 3.2. These *DER* resulting objective tests are represented by the bar graphs Figure 3.16, where the dark blue  for the clustering into 3 clusters and the light blue  for the clustering into 2 clusters.

The next step of the implementation is: The collection of the groups, the calculation of the PDF of each group, the finding of the variance of each group, configuration of the V-array, configuration of the Decision-array. The Decision-array is configured by the clustering of each row of the V-array alone into 2 clusters: high variances represent the mixture speech and low variances represent the dialogue speech. The outputs of this step are two arrays: the 1st array (the V-array) is used for the optimisation process by the choosing of the best groups and neglect the bad groups. The optimisation clears the road for the 2nd array, the Decision-array, by the clustering of good groups of this array of into 2 clusters, represent the semi-final segregation: the 1st is the dialogue speech mask and the 2nd is the mixture speech mask [70-72].

In addition to the optimized clustered Decision-array, the hierarchical clustering scenarios aid in creating the final version of the segregation binary masks to block the unwanted speech signal and permit the desired speech signal. The Final programing step is the writing of the output files of the above implementation on specific physical locations.





Briefly, the average *DER* for the 300 conversation between the 24 speakers, are tabulated in Table 3.3. They are plotted as a colour bars in Figure 3.16, where by the clustering into 3 clusters , by the clustering into 2 clusters , by the supposing of every 2-second switching instant  and by this chapter algorithm .

Table 3.2 The percentage average DER without the algorithm. The errors of the Mixture inside the dialogue E_{MISS} , the dialogue inside the Mixture E_{FA} and the overall errors E_{OVL} for the 55 FF conversations, the 105 MM conversations, and the 140 FM conversations (All = 300 conversations).

	By the clustering into 3 clusters			By the clustering into 2 clusters		
	E_{MISS}	E_{FA}	E_{OVL}	E_{MISS}	E_{FA}	E_{OVL}
F&F	34.0	36.7	35.1	22.3	16.9	20.1
M&M	30.9	41.4	35.1	23.1	18.7	21.3
F&M	38.2	28.3	34.2	30.4	15.2	24.3
All	34.9	34.4	34.7	26.4	16.7	22.5


The green bars  are for the processing when the switching instants are changing regularly each 2 minute. For those speakers, the average error of the mixture speech inside the dialogue speech and the dialogue speech inside the mixture speech is **11.6%**, more details are listed in the Table 3.3.

Table 3.3 The percentage average DER using the algorithm. The E_{MISS} , the E_{FA} and the E_{OVL} for 55 Female with Female F&F conversations, 105 Male with Male M&M conversations, and 140 Female with Male F&M conversations (All = 300 conversations).

	By the Clustering of the variances without the Optimisation			By the Clustering of the variances with the Optimisation		
	E_{MISS}	E_{FA}	E_{OVL}	E_{MISS}	E_{FA}	E_{OVL}
F&F	7.1	26.1	14.7	0.5	2.1	1.1
M&M	7.2	18.4	11.7	0.4	2.2	1.1
F&M	5.8	17.1	10.3	0.3	1.5	0.8
All	6.5	19.2	11.6	0.4	1.9	1.0

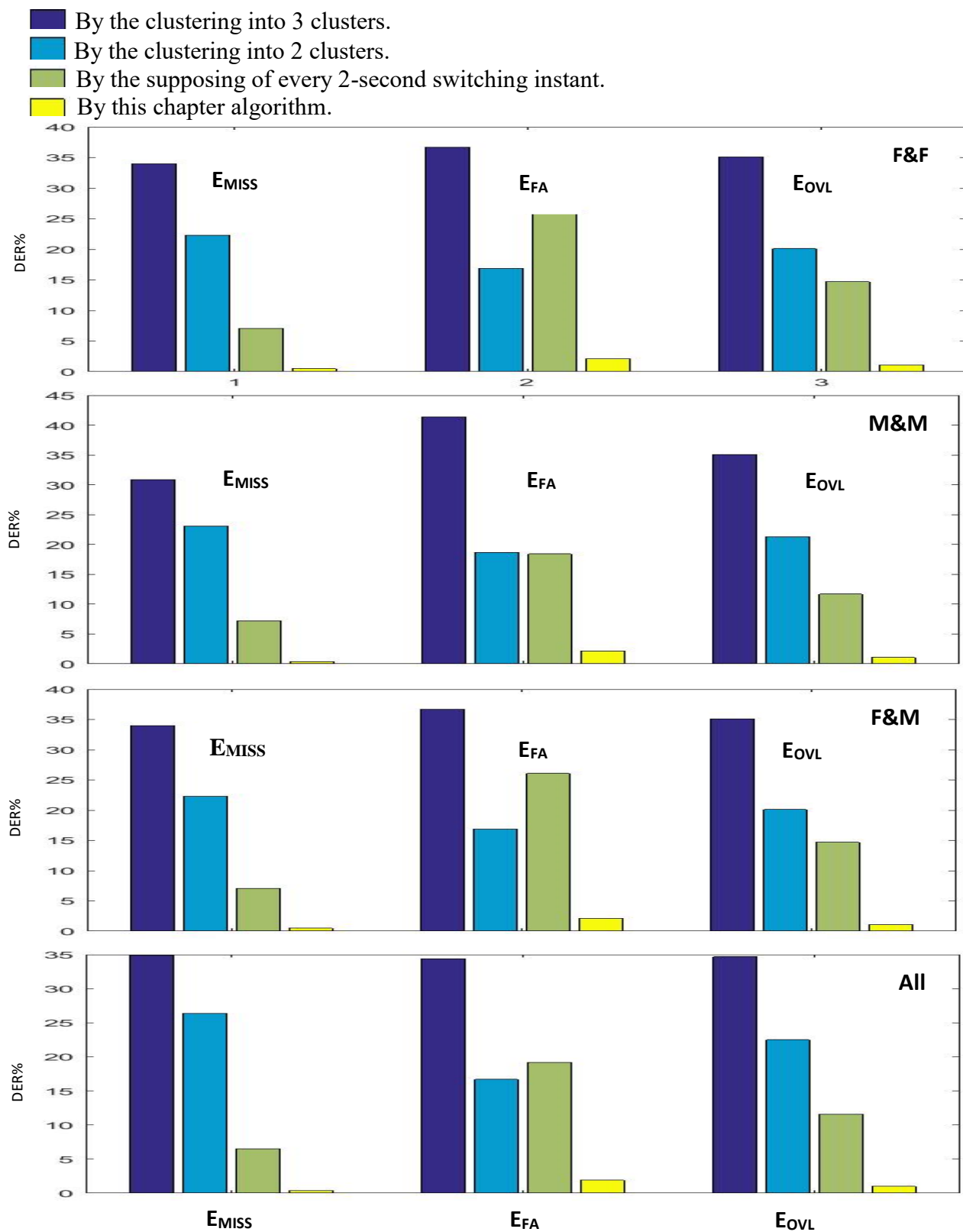


Figure 3.16 The percentage average DER of: The Mixture inside the dialogue (E_{MISS}), the dialogue inside the Mixture (E_{FA}) and the overall errors (E_{OVL}). They are for 55 FF conversations, 105 MM conversations and 140 FM conversations.

The following observations are drawn from the results of the 300 conversations' implementation:

- The average error of the mixture speech inside the dialogue speech and the dialogue speech inside the mixture speech is **1%**. Obviously, the average *DER* denotes that the algorithm is excellent and has an efficient performance.
- The maximum errors which are because of the spreading of a dialogue signal during the mixture speech period E_{FA} is 2.6% and the minimum is 0.3%.
- The maximum errors which are because of the spreading of a mixture signal during the dialogue speech period E_{MISS} is 1.4% and the minimum is 0.00001% $\approx 0\%$.
- For each conversation, the E_{FA} error is several times larger than the E_{MISS} error.
- The gender of the two speakers does not have any significant effects. Sometime the same-gender (MM or FF) conversation has fewer errors E_{FA} and E_{MISS} , than the different-gender (MF) conversation. Other times, the different-gender conversation has fewer errors than the same-gender.

From the conversations that have short periods of silence, the locations of the errors are at the switching instants (see Figure 3.15). If the conversation has periods of silence longer than the time of any neglected group, the errors could be inside the speech signal. For a short-time silence, the scenarios are helpful, but for a long-time silence, the scenarios might be harmful because they could spread to the non-sensible time error to be sensible. Figure 3.17 shows the input and the outputs of this case. In order to avoid such possibility, there are several algorithms which have the ability to remove or reduce the duration of the silence, such as the RAPT algorithm [8].

3.6 Comparison

The research algorithm is novel, but the novelty is not enough to evaluate its efficiency. To evaluate the algorithm rightly, its resulting objective tests have been compared with standard corpuses and recent literatures tests.

Summary of the comparison is illustrated and tabulated in Figure 3.18 and Table 3.4 together. The Missed-Speech Error Rate (E_{MISS}) is the important test to evaluate overlapped-speech detection. Obviously, that error rate for the algorithm is less than for most of standard corpus. According to the comparison between the research algorithm with collection of recent articles, the efficiency of the algorithm is better than most of these articles.

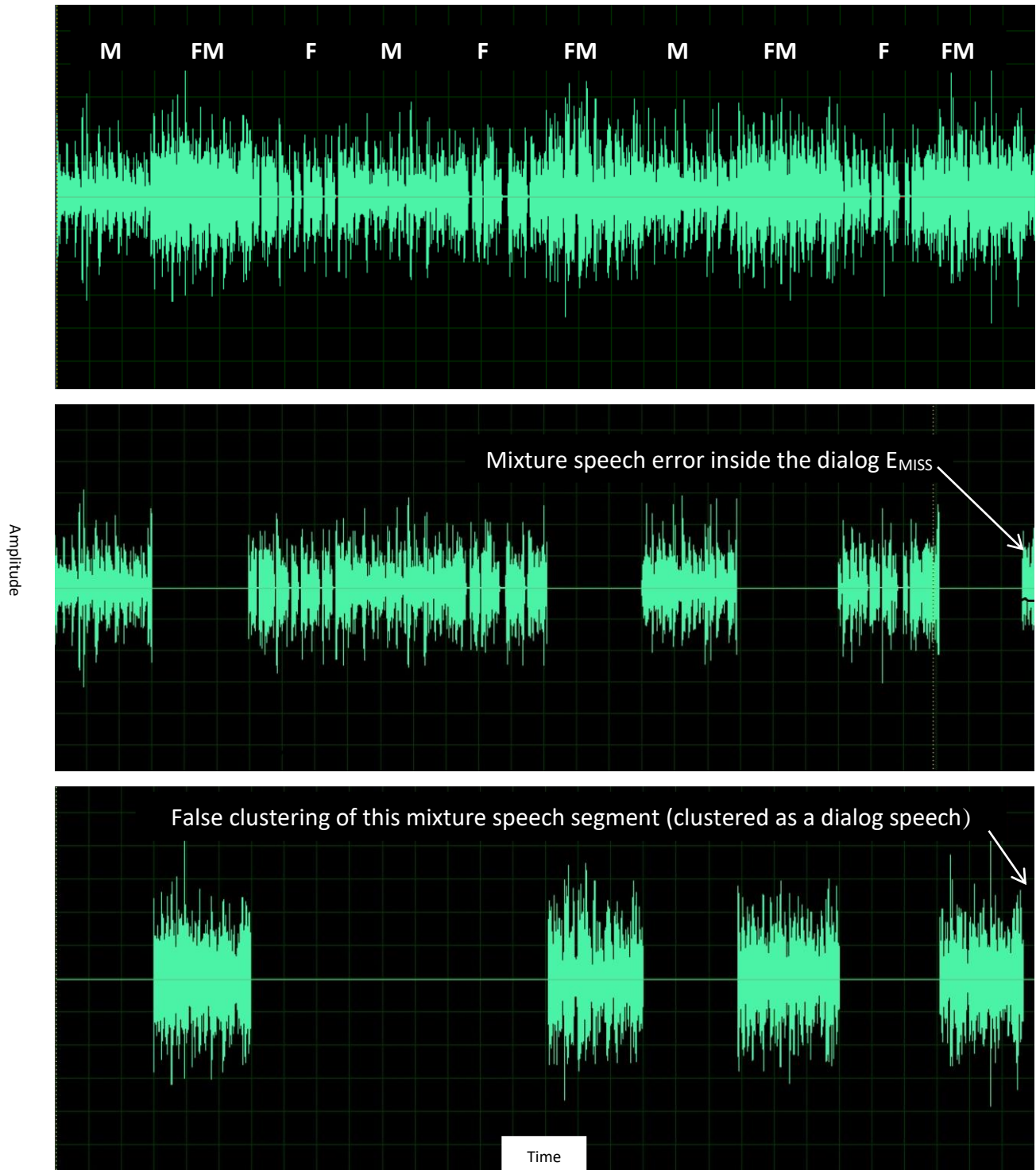


Figure 3.17 The implementation when the speech has a long period of silence during the mixture speech. The algorithm and the scenario suppose that the speaker is one because the other speaker is silent. There are horizontal-axes time-domain relationships between all the sketches.

The comparisons denote that the performance of this chapter algorithm is excellent. The subjective and objective test prove that. The comparison between the algorithm with the best standard professional speaker diarization corpuses, the algorithm could be categorized as professional overlapped-speech detection, because it has the same ability or better [70-72].

Table 3.4 comparison between the research algorithm with recent articles. The efficiency of the algorithm is better than most of these articles. N.B. F&F is Female with Female conversations, M&M is Male with Male and F&M is Female with male.

Researcher	Ref.	Test
K Boakye & et al.	[9]	DER = 0.07%, when the input conversation is high quality. DER = 0.04%, for the field-channel conversation
K Laskowski & et al.	[58]	0.36% relative error reduction
O Ben-Harush & et al.	[59]	0.6%, with 0.05% False-Alarm-Rate E_{FA}
H Pericás & et al.	[62]	0.74% improvement compared with the traditional algorithms.
R Yokoyama & et al.	[63]	About 0.26% improvement.
S Shum & et al.	[67]	0.81%
P Kenny & et al.	[68]	DER = 1.0%, for the summed-channel data for the Variational Bayes system. DER = 3.5% for Baseline system.
M-J Caraty & et al.	[69]	0.9%.
The research algorithm: H. A. Kadhim, L. Woo, & S. Dlay,	[70-72]	E_{MISS} = 0.5 for the F&F conversations, 0.4 for M&M, 0.3 for F&M and 0.4 for all the simulated conversations. E_{FA} = 2.1 for the F&F conversations, 2.2 for M&M, 1.5 for F&M and 1.9 for all the simulated conversations. E_{OVL} = 1.1 for the F&F conversations, 1.1 for M&M, 0.8 for F&M and 1.0 for all the simulated conversations.

The above table and the next table with graph illustrate clearly that the DER (E_{MISS}) of this chapter algorithm in the range of the standard corpuses. The algorithm is better than some of those corpuses for 2.2%. When the corpuses are better than the algorithm, the degradation of the algorithm is about 0.2%, i.e. negligible reduction.

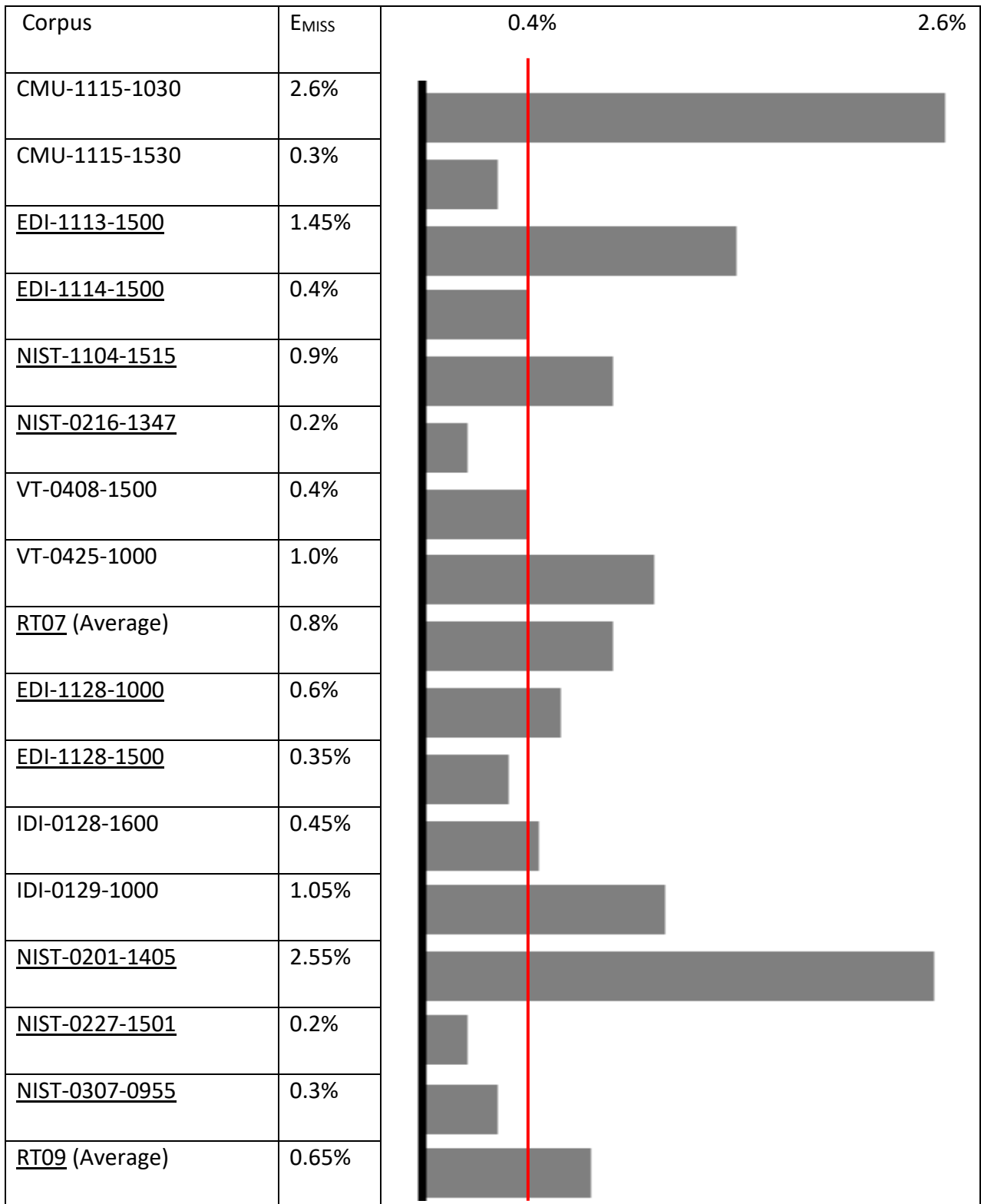


Figure 3.18 Missed-Speech Error Rate (E_{MISS}) comparison between the research algorithm with the standard speaker diarization corpora. The **RED-LINE** is the average E_{MISS} of this chapter algorithm. Obviously, the error rate of the algorithm is less than the error rate of most of these corpora [11]. The E_{MISS} % values are rounded to the nearest decimal point [70-72].

3.7 Summary

The observation signal of all the thesis system (spontaneous conversation of 2 speakers) inputs this chapter algorithm. The algorithm is a novel overlapped-speech detection algorithm. Excellently, the algorithm detects the switching instants from dialogue speech format to mixture, or vice-versa. Correct detection enables the algorithm to segregate those two speech formats, each one alone. The first segregated speech signal is the dialogue speech segments which is the input of speaker diarization process. In the Chapter 5, speaker diarization tool-box is invoked to isolate the speech of each speaker alone.

The second segregated speech signal is the mixture speech segments which is the input of speech separation process. In the Chapter 4 and the Chapter 5, speech separation separates the speech of each speaker alone.

The novel algorithm of this chapter has several steps to estimate the locations of the switching instances. RASTA-PLP is used to extract 13 feature for each frame of the input signal. The conventional clustering could not discriminate between the speaker and their mixture speech. Optimization loop improve that clustering. The optimization finds the best features and the worst. The features are capsulated inside group. The best group contains best features, and the worst group contains worst features. The worst groups are neglected and the best are adopted. The main factor to discriminate the mixture from the dialogue, is the variance of the PDF of each group. The fundamental group is 0.1-s, the second is 0.2-s and so on till the 32nd is the 3.2-s. The k-means clustering several times to draw the required borders between the clustered data.

For 300 simulated spontaneous conversations, 297 (99%) passes the subjective and the objective tests successfully. Only 3 (1%) failed, because they contain unacceptable periods of silent. Subjective tests denote that there are insensible errors at the switching time only. The error duration is fraction of second. The objective tests denote that the algorithm performance is equal the best current speaker diarization corpuses. The algorithm performance is better than most the traditional corpuses.

The following parameters have non-significant effects on the algorithm performance: genders of the speakers, sampling rate, resolution of samples, number of features per frame, subject of the conversation, length of the conversation, small periods of silence, the speakers of specific gender, the starting and the ending of the conversation, and the sequence of the mixture-dialogue speech formats.

The following parameters have significant effects on the algorithm performance: types of the extracted features must be RASTA-PLP for this algorithm, long period of silent cause misunderstanding for the algorithm, optimization step, the hierarchical clustering scenarios, and the basic conditions of the speech DSP (e.g. hopping period).

According to the correct output speech formats, the dialogue format segments are ready for the next speaker diarization process. Alone, the mixture format segments are ready for the next speech blind separation process in the Chapter 4. With outputs of the speaker diarization, the mixture format segments are ready for the next informed speech separation process in the Chapter 5.

Chapter 4. Blind Speech Separation by Filter-Bank, Non-negative Matrix Factorization and Speaker Clustering

4.1 Introduction

The algorithm of the Chapter 3 detects the swathing instants between each two adjacent different speech formats (i.e. the dialogue to the mixture or vice-versa). Those detections enable the research to segregate the main input spontaneous conversation into its two speech formats components. One of them is the mixture speech format. When the mixture speech is processed without any additive support of other information, the process is called blind speech separation (i.e. unsupervised machine learning system). In this chapter, the input signal is only the resulting mixture speech segments (segments), of the previous detection algorithm.

The algorithm of this chapter is a novel to separate the mixture speech signal, into its original separated speech of the speakers, independently each one alone. Each input segment of the mixture speech is processed entirely. The entire segment of the speech is important to discriminate between the parameters themselves, and the frequency domain components themselves. At first, the segment is framed by the standard overlapping-windows. Frame-by-frame, the input signal is passed through 65-subbands filter-bank analysis. Each sub-band admits, only to its dynamic range of the frequency domain components, to pass through it. Each sub-band output is the time domain signal of those components. The filter-bank outputs are 65 filtered signals. That filter-bank analysis enhances the separation capability of the next Non-negative Matrix factorization NMF technique. NMF alone has good ability to separate the mixture audio signals. That ability is limited for the composite mixture speech signal.

Each output signal of the filter-bank analysis is divided into 24 sub-signals by the NMF source separation. The total outputs of the NMF are the 65×24 sub-signals. The assistance of the filter-bank analysis to the NMF is not enough to separate the speech segment efficiently. To increase that efficiency, the mission of the filter-bank analysis and the NMF is only the separate of these sub-signals, but it is not the identity of the separated components of the sub-signals.

To identify the components of those sub-signal, speaker clustering method is used. Speaker clustering is the second phase of the speaker diarization. To perform it correctly, this phase needs the first phase of the speaker diarization, which is the speaker segmentation. To simplify the overall algorithm of this chapter, standard speech frame is contributed as a segment, for the speaker segmentation process. Reliable exist speaker clustering toolbox is invoked to identify the standard

frames of the separated sub-signals.

The output clustered standard audio frames are masked to share them between the two speakers. Soft mask is better than the binary in this chapter to share each speaker components, and avoid unwanted components. To produce the final separated speech signal for each speaker, filter-bank synthesis is used to collect these shared components.

Subjective and objective tests of the simulates segments of the chapter 1 conversations, denote that the algorithm is very good compared with the recent well-known literatures [22, 87].

4.2 Source Separation

In most cases, input observation signals of DSP jobs are: mixed-information signals and added-noise signals. The mixed-information signal contains various components, e.g. an audio signal contains a music signal with a voice of singer speech signal. Those mixture signals, sometimes are homogeneous signals, but other times, are non-homogeneous signals. Splitting of those different signals into their original components is important job for the DSP researchers. In addition to that, these signals suffer from the time domain additive-noise. Removing of such unwanted noise is an important mission that faces the researchers. The mission could be expressed as a splitting methodology to extract the wanted signal without (or with as little as possible of) the unwanted noise signal. Generally, the splitting methodology is called Source Separation. For speech observation signal, such speech-DSP process is called Speech Separation. According to the speech-DSP researches achievements of that separation field, the source separation of audio signals is less complex than speech signals. The complexity is due to the physical similarity among parameters and characteristics of the speech and the speakers. As well as, the output separated signals have common characteristics among them. The similarity is less in the audio case, where the separated signals are a speech which mixed with: sounds of machine, sounds of music instruments or with other sound [148].

In the speech separation, there are n speakers those take part in a conversation. The speakers are speaking simultaneously for a specific period. The individual speech signals of those n speakers are x_1, x_2, \dots, x_n . In speech separation, these signals are called “targeted-speech signal”. The resulting mixture signal x_{mix} of the conversation session is the observation signal of the speech separation processing. The knowhow of the mixing (which produces that input signal) is important for the simplicity or the complexity of the method for the separation process. When x_{mix} is produced

by the algebraic sum of the weighted x s, the system is called Single Channel Speech Separation. In this thesis, there are the following two speakers: Female F with Male M (the mixture signal is symbolled by FM), Female F with Female F (the mixture signal is symbolled by FF), or Male M with Male M (the mixture signal is symbolled by MM). Generally, the first case FM (fm in time domain) is used as an example for the describing details about specific process.

$$x_{mix} = \sum_{i=1}^n a_i x_i \quad (4-1)$$

$$x_{mix} = fm = a_f f + a_m m \quad (4-2)$$

$$X_{mix} = FM = a_f F + a_m M \quad (4-3)$$

For r -channels speech separation, there are r observation signals. Each one of them (x_{mix}^s) is produced by its specific weighted algebraic sum of them, i.e.:

$$x_{mix}^s = \sum_{i=1}^n a_{si} x_{si} \quad (4-4)$$

where s is 1 to r (r is less than or equal n). There are r transducers (e.g. microphones) for collecting the r observation signals.

For the previous several decades, audio and speech-DSP researchers were enrolling with the separation challenge and how they could overcome it by recovering the original targeted-speech of each speaker individually. The most achievable separation techniques are: The Principal Component Analysis PCA [149], the Independent Component Analysis ICA [150], the Non-negative Matrix Factorization NMF [79] and the Computational Auditory Scene Analysis CASA [151]. Speech separation stills the most challenge area which poses the audio and the speech researchers. Relatively, the single channel speech separation has the most challenge, because it has fewest helpful attributes to solve its dilemmas.

Originally, the speech separation is called “The Cocktail-Party Problem” because it is dealing with the mixture speech of multi-speakers who are speaking simultaneously. On the other hand, speaker diarization is called “Who Spoke When?”, because it is dealing with the dialogue speech of multi-

speakers. The dialogue is a conversation format, where only one speaker is speaking at any specific time during the conversation, but other speakers are silent [11].

For the research thesis, the separation process is treated by two different approaches. The first is this chapter approach, where the separation process is not supported by any other information rather than the input observation signal (Blind Speech Separation) [152]. The approach uses the four-combinational well-known techniques: Filter-Bank Analysis, NMF, Speaker Clustering and Filter-Bank Synthesis.

4.3 Functional Block Diagrams and Waveforms

Continuing with the details of chapter 1 and chapter 3, this chapter process is the second block of that overall system (see the Figure 4.1). The bold block/Figure 4.1 contains the two parts (sub-blocks) of this chapter algorithms. The first sub-block is a blind speech separation which separates the mixture speech signal. The second sub-block is the speaker clustering which assigns each segment to its speaker. This sub-block is not involved under the research thesis, i.e. an exist speaker diarization application has been used for the assignment task. The separation sub-block, only separates the mixture signal but cannot identify the output separated segments [22, 87].

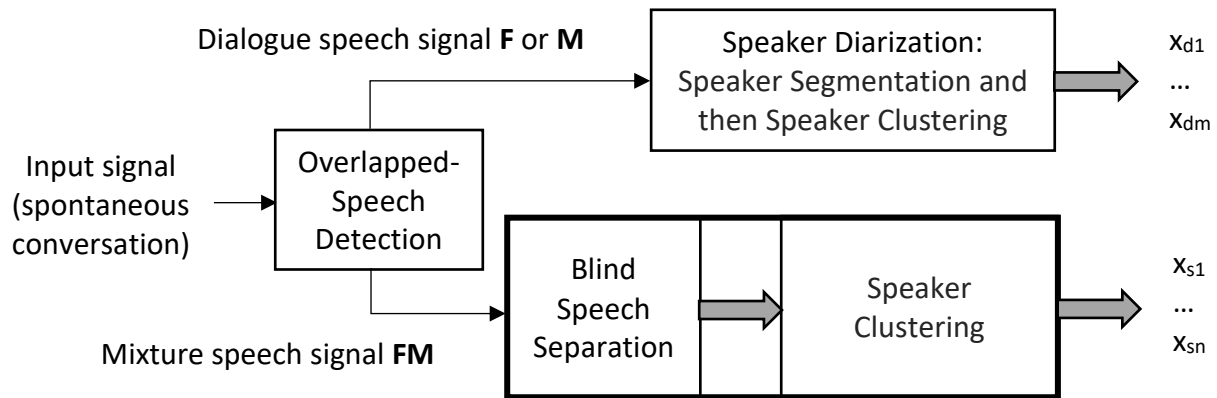


Figure 4.1 Chapter 4 overall system. The input is spontaneous conversation signal and the outputs are the individual speech signal of all the speakers. The system is unsupervised Machine Learning system.

Suppose there are N_s processed segments. Number of speakers per the i^{th} segment is N_{spi} . Number of the total output segments N_{ts} are:

$$N_{ts} = \sum_{i=1}^{N_s} N_{spi} \quad (4-5)$$

When number of the speakers is two, N_{ts} is two times N_s , because there are two same speakers who are talking during each segment. There are few complexities when the speakers are three. The complexities are a result of the unknown number of speakers per each segment. They are two speakers sometimes and three speakers at other times.

The process becomes more complicated for the case of four speakers, because the speakers per segment, probably two, three or four. The deduction is: The processing complexity depends on the number of speakers in the conversation. Number of the output segments of the 1st sub-block is the summation of number of speakers/1st segment plus number of speakers/2nd segment, and so for all the mixture speech segments (FM) during the conversation.

By the continuing with the Chapters 1 and the Chapter 3, the overall system input signal is a spontaneous conversation speech, the (a)/Figure 4.2. Sometimes, it is a single speaker signal, i.e. dialogue speech (either F or M is speaking). Other times, it is a multi-speakers signal, i.e. mixture speech (both F & M are speaking together simultaneously). This FM mixture speech is the overlapped-speech during the conversation, the (b)/Figure 4.2. The FM is the input of this chapter block. The outputs of this chapter functional block are the separated speech signals of F alone (the (c)/Figure 4.2) and M alone (the (d)/Figure 4.2). Typically, the input of this chapter algorithm is the (b)/Figure 4.2 and the outputs of this chapter algorithm is the (c) and the (d)/Figure 4.2.

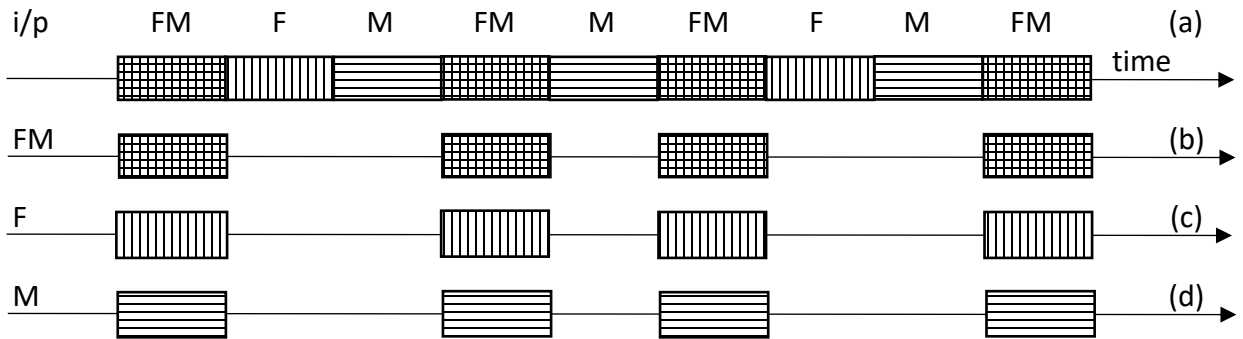


Figure 4.2 Arbitrary spontaneous conversation, the dialogue between Female F (vertically-lined) alone and Male M (horizontally-lined) alone. the mixture FM is both simultaneously (cross-lined). Single Channel Speech Separation Procedure. There are horizontal-axes time-domain relationships between all the sketches.

Briefly, this Chapter speech separation is adapted by a concatenated implementation of: Filter-bank analysis technique (the 1st block/Figure 4.3), NMF technique (the 2nd block/Figure 4.3), speaker identification (the 3rd block/Figure 4.3) and then filter-bank synthesis (the 4th block/Figure 4.3). One output of the Chapter 3 are segments of an overlapped-speech; they are the input of the first block. The recovered speech is the output of the clustering block. The Figure 4.1, the Figure 4.2 and the Figure 4.3 are the detailing functional block diagram of this Chapter system. Generally, the Figure 4.5 is the flowchart that concepts the behavior of those blocks. The mixture signal, sequentially passes through those four main processing blocks: filter-bank analysis, NMF, speaker clustering and then filter-bank synthesis. The speaker clustering is the last phase of the speaker diarization process. An exist tool-box is invoked to implement the speaker clustering phase.

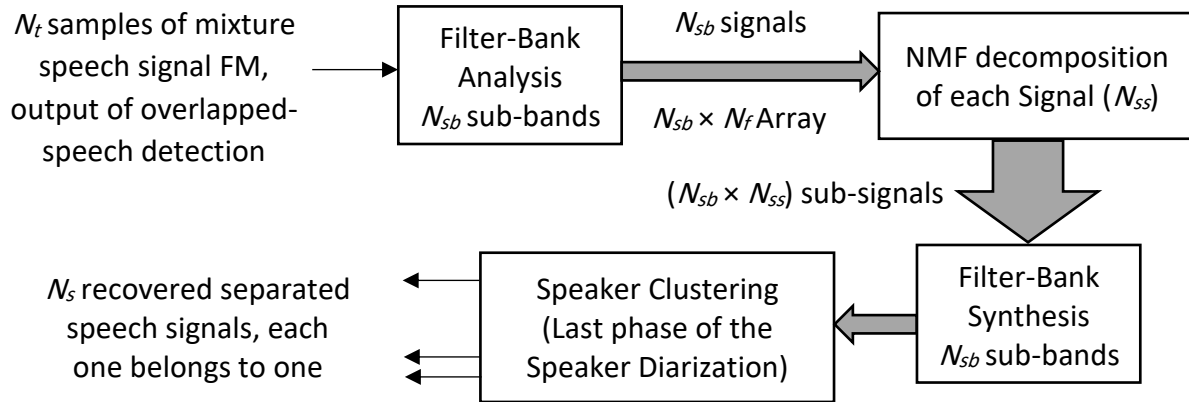


Figure 4.3 Functional block diagram of this chapter algorithm.

4.3.1 Preparation of the Required Resources

The input speech signal of the algorithm is a mixture speech which is one of two outputs of the overlapped-speech detection algorithm of the Chapter 3. The mixture is a composite signal of multi-speakers those are speaking simultaneously at each segment. The mixture speech signal of the overlapped-speech detection algorithm is non-adjacent segments which are extracted from the original spontaneous conversation. In this chapter, the segments are processed segment-by-segment. Processing rules of each segment, are the same rules for other segments, although the parameters and characteristics of each segment are different compared with other segments.

Initially, the possible conditions, the genders, the resolutions and the sampling rates should be considered. For each gender, enough number of speakers are chosen for the conversations. The

duration of each processed segment is 30 s. The 30-s period is a compromise choice, because the period of more than 30 s does not provide an extra significant information. The period of less than 30 s, may be is not enough to cover most possibilities of the conversations.

The required time to implement this chapter algorithm is the conclusive factor to pick up the suitable choices and to avoid the non-suitable. The 16000 sample/s sampling rate, compared with the 8000 samples/s increases, rapidly the implementation time, and does not provide extra useful results. For this purpose, the 8000 sample/s sampling rate is chosen. Similarly, the 16 bit/sample resolution instead of the 32 bit/sample is used. Arbitrary, 10 females and 14 males are the speakers of the input conversations. Number of the possible conversations for FF is 55, for MM is 105 and for FM is 140. The total number of these conversations is 300, but the required time for experimenting these conversations are so long and not feasible, so only 51 essential conversations have been simulated and then tested.

4.3.2 *Filter-Bank Analysis Technique*

Filter-bank is an all-pass frequency domain filter technique. Originally, it is an analog technique, then it is modified to adapt with the digital technology. The main application of the digital filter-bank was the compression of speech and audio signals. For the compression application, the partial decimation of the input signal samples (of the shifted sub-bands) is conducted by down-sampling its sampling rate. The interpolation numerical analysis is used to up-sampling (of the reverse shifted sub-bands) the sampling rate of the output signal. This is the main concept of the so-called the Vocoder.

Each filter-bank has N_{sb} sub-bands filters. The first filter is a low-pass filter and the last filter is a high-pass filter. The other $(N_{sb} - 2)$ filters are band-pass filters. The dynamic-range of each filter resides in its corresponding sub-band. According to Discrete Fourier Transform DFT, in time domain each frame of speech signal consists of N_w samples. In the frequency domain, each i^{th} (where i is 0 to $(N_w - 1)$) sub-band of that frame is a mirror-conjugate of the $(N_w - i)$ sub-band, except the first and the center sub-bands. For this reason, the first half plus the center sub-bands are considered for the spectral description of the speech signal (i.e. $1 + (N_w/2)$ sub-bands). To increase the resolution of the filter bank and the transformed speech signal, number of the sub-bands should be increased, i.e. by increasing the number of speech signal samples per frame in the time domain. On the other hand, there is limitations against the increasing or the decreasing of the

time domain period of the processed speech frames. For the efficient speech-DSP processing, the period should be inside the range 8 ms to 20 msec. This time domain range is corresponding to the range of 50 Hz to 125 Hz in the frequency domain. In most cases, processing the sub-band inside this range is not efficient till the best limit (the 50 Hz resolution) is used. To increase the resolution (decrease the sub-band width), the successful alternative is the wider overlapping-windows of speech frames. The range of each overlapping-window of speech frame is 16 ms to 40 ms. The relevant frequency range of that alternative is the 25 Hz to 50 Hz. These ranges are qualified when the non-overlapping between each two adjacent frames is inside the constrain of 8 ms to 16 ms range. The overlapping-window time equals subtracting the non-overlapping time from the frame time. The non-overlapping-window period, is called the hopping period. That alternative is a compromise solution which increases the resolution (decrease the sub-band width) under the proper constrains of the speech and the audio DSP. In the case of 8000 sample/s sampling rate, the range of the overlapped period is 128 to 240 samples of each frame with its adjacent frames. The frame of 240 sample, could be transformed to its frequency domain using the slow DFT, but the best transformation is by using the fast FFT. For the FFT transformation, number of input samples per frame must be Radix-2, so the nearest approximated number to 240 is 256. The hopping period is 64 to 128 samples each hopping from any frame to its next adjacent frame. In the case of 16000 sample/s sampling rate, the range of the overlapped frames is 256 samples to 480 samples each frame. The nearest approximated radix-2 number to 480 is 512 samples per frame. The hopping period is 128 to 256 samples each hope from any frame to its next adjacent frame. Although the width of any processed frame is 16 ms to 40 ms, the effective period is about 8 ms to 20 ms. The effectiveness is due to impact of the window scaling. Bilaterally and gradually, the window attenuates the amplitude of the input samples. The scaling is conducted by the element-wise multiplication of the frame samples by the element values of the window vector. There are many well-known windows e.g. Hamming, Hanning and Kaiser.

The filter-bank has two stages: the analysis stage and the synthesis stage; Figure 4.4 the input speech signal is the inputs of all the analysis blocks of the filter bank. The output of each analysis sub-band is the frequency domain content of the speech signal for that sub-band. The outputs of the analysis stage could be passed through a desired DSP processing block. The processing is implemented sub-band by sub-band (sequentially) or sub-band with other sub-band(s) (parallel processing). The outputs of that block are the input of the synthesis sub-bands blocks. The algebraic

summation of the outputs of the blocks is the main output of the filter bank. Although the filter-bank is a frequency domain methodology, in most cases the calculations of the general filter-bank and a specific filter-bank are implemented in the time-domain. The reason for that, is the consumed time for the required calculations of those general and specific filter-banks [22, 87].

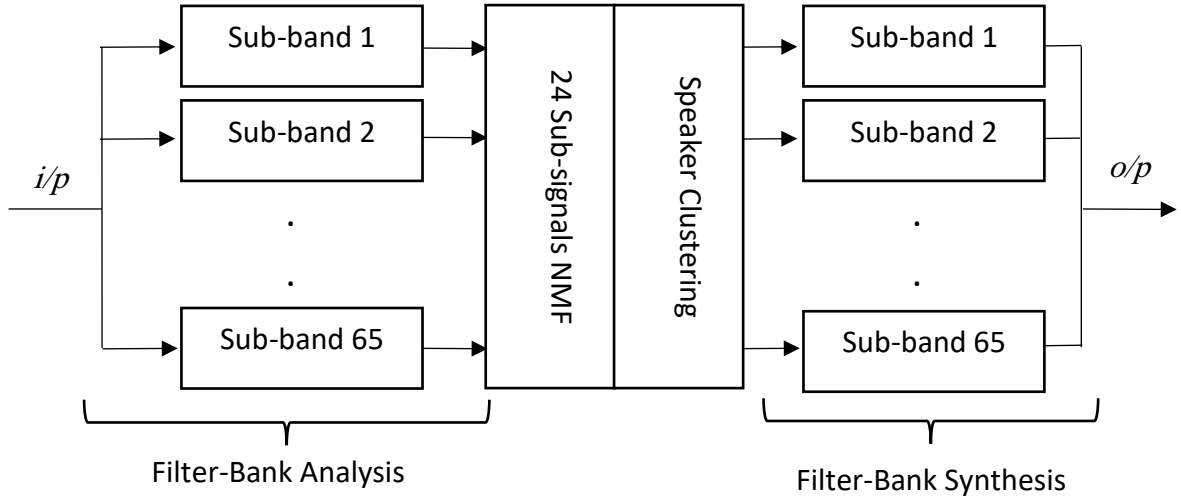


Figure 4.4 Block diagram of this chapter algorithm. Input signal (i/p) is the mixture speech, and Output signals (o/p) are the separated speech.

4.3.3 Non-negative Matrix Factorization NMF

Mathematically, matrix is a powerful arrangement for the data in both number and function forms. The matrix, the determinant and the vector are the terms of the linear algebra. In the computer programming field, the equivalent term of any one of them is “array”. The vector is a one-dimensional array. The determinant is a square two-dimensional array. The matrix is a multi-dimensional rectangular array. The array is an adequate container for tabulating the information and data, because the fact that these data should be manipulated by a machine (e.g. computer).

These data reside their meaningful locations inside specific matrices. The data are facing a problem of the dramatically increasing of these data. The increasing causes a huge expansion in the capacity of the storage devices those save these data. These increased-data cause more manipulating time also. There are requests to create reduced data which equivalent to these original huge data. To find the equivalent, factorizing technique could be used for that purpose.

Instead of the original main huge matrix $[S]$, the factorization transforms it to reduced multi-matrices. By the factorizing:

$$[S] \approx [W][H] \quad (4-6)$$

$$[S] \in \mathbb{R}^{r \times ss, \geq 0}, [W] \in \mathbb{R}^{r \times ss, \geq 0}, [H] \in \mathbb{R}^{ss \times c, \geq 0}$$

where, the data matrix $[S]$ has r rows and c columns, $[W]$ has r rows and ss columns, and $[H]$ has ss rows and c columns.

$$[e] = [S] - [W][H] \quad (4-7)$$

where, $[e]$ is the error matrix. The factorization algorithms are based on the feedback iteration programming. The calculated error norm $\| [e] \|$ determines the divergence condition and the accepted tolerance, to finish the machine running and accept the approximated factors $[W]$ & $[H]$.

$$\min([W] \text{ or } [H]) \sum_{i=1:r, j=1:c} |[S] - [W][H]|^2 \quad (4-8)$$

$$[S] \in \mathbb{R}^{r \times ss, \geq 0}, [W] \in \mathbb{R}^{r \times ss, \geq 0}, [H] \in \mathbb{R}^{ss \times c, \geq 0}$$

For the capacity of the required storage, $[S]$ resides $(r \times c)$ memory locations, and $[W]$ plus $[H]$ reside $(r \times ss) + (ss \times c)$ memory locations. The difference between the required locations for the original $[S]$ matrix and the required locations for the factors $[W]$, and $[H]$ matrices depends on the difference between ss and the minimal value of r or c . This storage reduction, maybe million times, i.e. the required storage for $[S]$ is million time the required storage for $[W] + [H]$. The condition for using the Non-negative Matrix Factorization NMF is the fact that all the elements of $[S]$ must be non-Negative values (i.e. ≥ 0). The resulting $[W]$ and $[H]$ have non-Negative values elements also. To achieve equation (4-8), the following well-known algorithms and methods are used:

- Lee algorithm [74, 75].
- Brunet NMF algorithm [153].
- KL-NMF algorithm [154].
- Frobenius-Norm NMF [155].
- The Offset NMF method.
- The ns-NMF, the ls-NMF, the pe-NMF and the si-NMF algorithms.
- The SNMF/R and the SNMF/L [156].

For Lee algorithm, **Appendix D** has been added to the thesis to detail the description of NMF.

To explain how the NMF has been exploited for the speech separation job, the following describes the main idea for that. The total duration of the mixture speech segment is T_t , the duration of the overlapping-window speech frame is T_w and the duration of the hopping is T_h . Their corresponding numbers of samples per each frame are N_t , N_w and N_h respectively, where:

$$N_t = f_s \times T_t; \quad N_w = f_s \times T_w \quad \text{and} \quad N_h = f_s \times T_h \quad (4-9)$$

The total number of the processed frames are:

$$N_f \approx \text{floor}\left(\frac{T_t}{T_h}\right) \approx \text{floor}\left(\frac{N_t}{N_h}\right) \quad (4-10)$$

where, $\text{floor}(\cdot)$ is the down-rounding floor function which approximates its argument to the nearest less integer. Since N_w is number of the samples per each input frame in the time domain, number of the frequency domain sub-bands is $1+(N_w/2)$. The FFT for all the frames of the segment speech could be arrange as the $(1+(N_w/2)) \times N_f$ spectrum matrix $[S]$. In the speech-DSP, due to the ignorable effect of the phase variations, the frequency domain description is the absolute values of the spectrum magnitude. The elements of $[S]$ are positive values those represent the magnitude of their sub-bands. According to this property of $[S]$ matrix, NMF is applicable on $[S]$, i.e. could be factorized into two factorizing matrices: $[W]$ and $[H]$. Any i^{th} row of $[S]$, is the i^{th} vector $[D_i]$ which contain the spectral analysis of the i^{th} sub-band. Any j^{th} element of the $[D_i]$ vector is:

$$S_{ij} = \sum_{n=1}^{1+N_f/2} W_{in} \times H_{ni} \quad (4-11)$$

where W_{in} is the i^{th} row- n^{th} column element of $[W]$ and H_{ni} is the n^{th} row- i^{th} column element of $[H]$. According to Equation (4-11), each sub-band of each frame is the summation of the multiplication results of its spectral base by the activation weights of each sub-band. The full spectrum (i.e. all the sub-bands) of that frame is the concatenation arrangement of all these sub-bands. For that full spectrum, $[W]$ is the Spectral-Basis matrix of the filter-bank analysis and $[H]$ is the Activation-Weights matrix of the filter-bank analysis. Obviously, thus matrix manipulation has the ability for increasing the resolution of the frequency domain analysis. It seems that the

resulting ss number of sub of the sub-bands could produce ss number of their corresponding sub-waveforms in the time domain. This waveform generation has a splitting action for the original waveform of each sub-band, i.e. it has the ability for the separation of each sub-band components. Generally, the NMF technique is used, widely for the audio separation depending on the above spectral analysis of the mixture audio signal. In this Chapter, filter-bank analysis has been used for sustain the ability of the NMF for the separation of a mixture speech signal. Although the NMF has good ability for the separation of a mixture audio signal, NMF has poor capability for the separation of the mixture speech signal. Instead of the mixture speech signal itself, in this Chapter the NMF has applied on all the sub-bands signal, sub-band by sub-band. Since the NMF can split its input signal into ss signals/sub-bands and there are $1+(N_f/2)$ sub-bands, the total number of split signals are $ss \times (1+(N_f/2))$.

These several hundreds of signals are split but their speakers' identification have a lot of errors. The errors are since each waveform signal belongs to the multi-speakers. Instead of all-the-signal identification, the process deals with the frames one-by-one, inside each waveform signal. This Identification could be executed, successfully by using the speaker clustering algorithms. The speaker clustering is the second phase of the speaker diarization process. To implement the clustering in this chapter, an existing reliable speaker diarization toolbox have been used. The toolbox is an open-source package which available in the [GitHub](#) institute website. Since each conversation has N_f frames, the total number of the frames are $(ss \times N_f \times (1+(N_w/2)))$, Figure 4.5. The identified sub-frames are summed to produce the desired speech signal of a specific speaker. The other unwanted signals are masked for this speaker but they are considered for the other speakers. There are two types of the masks: the binary and the soft masks. The binary mask belongs all the specific signal for a specific speaker, and nulls the other speakers. The soft mask shares the signal among all the speakers. The sharing is done according to the distances of the signal parameters from their references [22, 87].

4.3.4 *Speaker Clustering*

Speaker Diarization is the speech-DSP which deals with dialogue speech format of multi speakers. The format of this speech is produced by speaking of only one speaker during the multi-speaker conversation. The diarization DSP consists of two phases: The Speaker Segmentation and the Speaker Clustering. Sometimes, a mathematical model (e.g. HMM or GMM) could support those

stages. For the speaker clustering, the Hierarchical Clustering Scenarios should be requested to reduce the probable clustering weak points. There are many successful applications to implement these phase of the speaker diarization [157, 158]. For this chapter algorithm, an existing application (toolbox) has been used for clustering the segments (the frames) of the separated speech.

For each sub-band, the NMF separates its *ss* sub-signal. These sub-signals are separated, but they are not assigned (identified). The speaker clustering is the second phase which needs the first stage “the speaker segmentation” [159]. In this chapter algorithm, the frames partitioning is used instead of the segments partitioning.

In this chapter algorithm, the binary mask and the soft mask has been used. The soft mask is calculated with respects to two distances. They are the amplitude distance and the energy distance. The amplitude distance is calculated for the amplitudes of the signals and the amplitude of the references. The energy distance is measured by the squaring of their amplitudes (i.e. the power of the signals and the power of the references). For each sub-band, the NMF separates its *ss* sub-signal. These sub-signals are separated, but they are not assigned (identified). The speaker clustering is the second phase which needs the first stage “the speaker segmentation” [159]. In this chapter algorithm, the frames partitioning is used instead of the segments partitioning.

The Figure 4.5 illustrates the flowchart of the general description of this algorithm Chapter.

4.4 Experiments

The above algorithm is experimented by the using of the software environments: MATLAB, Audacity, notepad++ ... etc. The main problem which faces the implementation is the long time to complete the processing of each mixture speech segment. For that, many selections are chosen with taking into account of this long-time effect. For each conversation, number of speakers is two persons. The speech of 55 Female with Female (FF), the speech of 105 for Male with Male (MM) and the speech of 140 Female with Male (FM) should be prepared because number of the speakers are 10 Females and 14 Males. The machine running time to complete such large number of conversations is unpractical. The accepted alternative implementation is the arbitrary choosing of 17 FF, 17 MM and 17 FM (i.e. total number of the conversations is $17+17+17=51$). To reduce that running time, the sampling frequency is 8000 sample/s with 16 bit/sample resolution. Each segment of the conversation is 30 s of mixture speech. Both speakers are talking simultaneously during all the 30-s period. The conversation segment consists of $30 \times 8000 = 240,000$ samples (N_i).

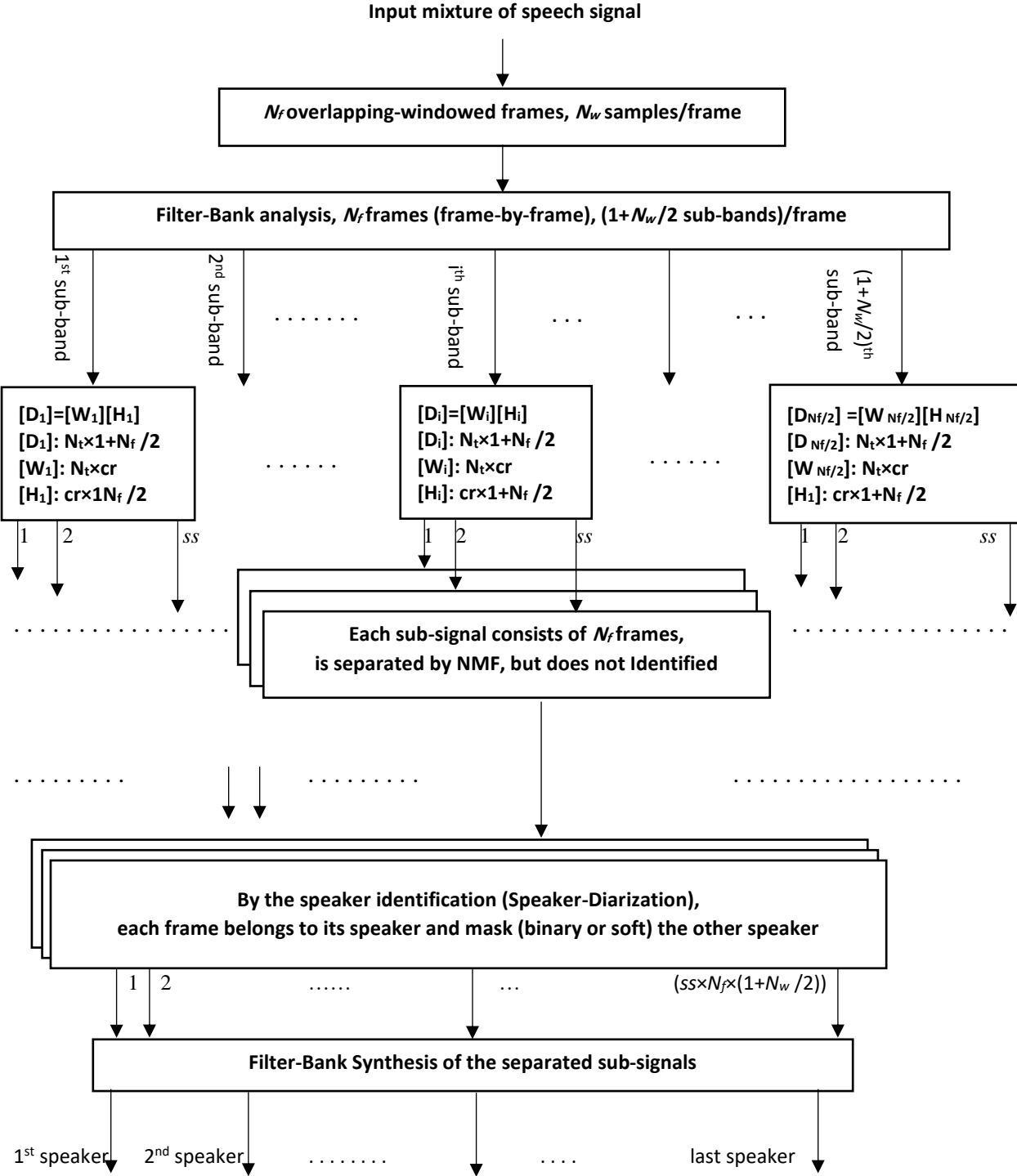


Figure 4.5 Flowchart of Chapter 4 algorithm. The input mixture speech signal is divided into N_f frames. Each frame is analysed to $(1+(N_f/2))$ sub-bands. Each sub-band signal is separated into ss sub-signals. Each sub-signal is divided into N_f frames. Each frame is assigned to its speaker; then synthesised of the wanted signals and masking the unwanted signals.

The overlapping-window frame of speech is 16 ms, i.e. $0.016 \times 8000 = 128$ samples (N_f). The hopping time is 8 ms, i.e. $0.008 \times 8000 = 64$ samples (N_h). By using the Equation (4-10), the approximated total number of the processed frames (N_f) is 3750 frames. The conversations should be stored as *FileName.wav* audio files in a computer machine.

After the above preparing and partitioning of the speech conversation is the first step by retrieving the .wav file from its physical location in a specific computer machine. Since each frame consists of 128 samples in the time domain, its transformation vector in the frequency domain consist of 128 complex-number points (The MATLAB code is appended).

Since the second half of X is the mirror-conjugate of the first half, (half +1) is enough to describe its frequency domain analysis, so, number of the regarded sub-bands is $(N_w/2) + 1 = 65$ sub-bands. The IFFT of each sub-band of X with its mirror-conjugate of X also, produces the speech signal of that sub-band. This overall FFT (Spectrogram) and the IFFT of a specific sub-band is the equivalent to the Finite-Impulse-Response FIR band-pass filter of that sub-band.

Instead of the traditional time domain calculations of FIR, the process is done inside the frequency domain. Now, the original segment of mixture speech signal is filtered to 64 mixture of speech sub-signals. Up to these resulting waveforms, the analysis part of the filter bank has completed its task. The first sub-band (the low-pass filter), several of the second to the sixty-third sub-band (the band-pass filters) and the sixty-fourth sub-band (the high-pass filter) are illustrated in the Figure 4.6. The speech signal FM is a mixture of Female F and Male M of the TIMIT standard library [20].

The next step is the NMF speech separation for the 64 sub-signals one-by-one. The following NMF process are similar for all these sub-bands, so the calculations for the first sub-band is like the 2nd sub-band process, like the 3rd sub-band process, and so on. The calculations are similar from the 1st to the 64th sub-band process. To perform the NMF separation, the absolute-magnitude values of each sub-signal spectrogram should be calculated. Number of $[W]$ columns = Number of $[H]$ rows = Number of the output sub-signals per each sub-band = ss . Due to the required running time, the minimal accepted ss is about 24 for 65 sub-bands and 8000 samples/s sampling rate. The total output sub-signals are $65 \times 24 = 1560$. Arbitrary, Figure 4.7 shows the 24 output sub-signals of the 3rd sub-band of a mixture speech conversation between F and M of the TIMIT standard audio library.

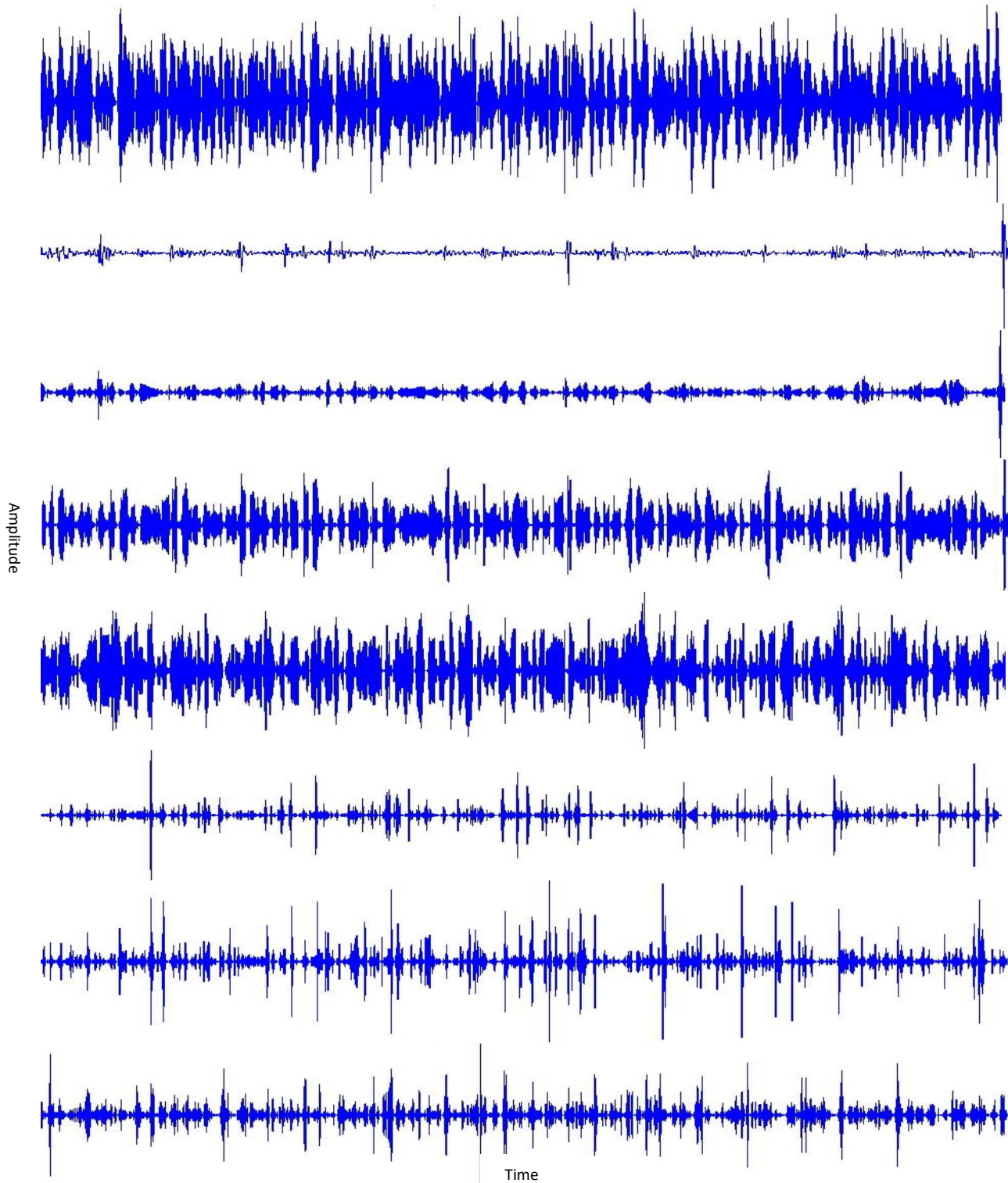


Figure 4.6 Specimen of the filter-bank analysis. The (a) is the FM mixture speech signal. The (b) is the 1st sub-band (Low-Pass filter), the (c) is the 2nd sub-band, the (d) is the 3rd sub-band, the (e) is 4th sub-band, the (f) is the 32nd sub-band, the (g) is the 33rd sub-band and the (h) is the last high-pass filter (the 64th sub-band). There are horizontal-axes time-domain relationships between all the sketches.

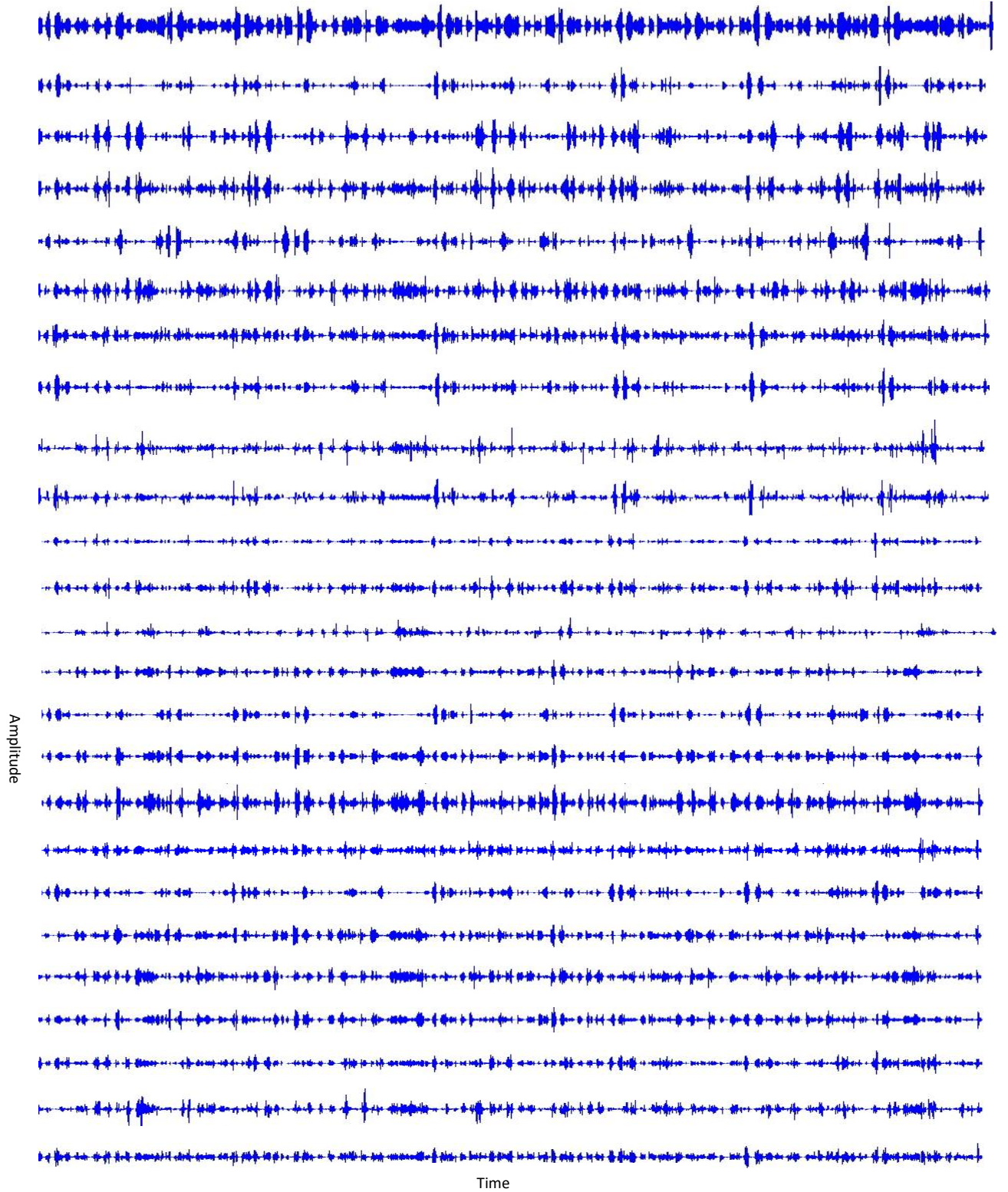


Figure 4.7 Specimen of the NMF speech separation. Number of the separated signals is 24 sub-signals. The 1st is the input, which is the 3rd sub-band (Figure 4.6). The 2nd to the 24th are the output sub-signals. There are horizontal-axes time-domain relationships between all the sketches.

After the testing of each sub-signal alone, the results appear that each sub-signal is a mixture signal of the F and M speakers. The mixing situation of these sub-signals is less complexity than the mixing situation of the input sub-band signal and the original mixture observation signal. These results are expected due to limited ability of the NMF separation for the speech signals [22, 87].

Since the improvement is useful but not enough for the accepted speech separation, two modifications are proposed to reach the accepted results. The first proposal is the only separation process because it have partial ability for that. The process separates components of the sub-signal, but do not identify for which person those components of the sub-signals belong. The speaker clustering can identify to which speaker the sub-signals (segments) belong. According to the speaker diarization principles, initially the speaker segmentation should prepare the suitable segments before that. To simplify the speaker segmentation job, the speech frames is expressed as the segments. It is the second proposal to improve the NMF technique for achieving the separation.

4.5 Result and Test

The above algorithm is implemented step by step for the chosen 34 speakers. They are 17 female speakers (F1 to F17) and 17 male speakers (M1 to M17). Arbitrary, the Female-with-Female mixture speech are: F1-F2, F2-F3,, F16-F17 and F17-F1. The Male-with-Male mixture speech are: M1-M2, M2-M3,, M16-M17 and M17-M1. The Female-with-Male mixture speech are: F1-M1, F2-M2,, F16-M16 and F17-M17 (one of them is the mixture speech of F and M of the TIMIT standard library, which is called FM).

The FM mixture speech is the simultaneously conversation between the female F and the male M from the standard audio and speech TIMIT library. The duration of that mixture speech is 30 second (the (a)/Figure 4.6, the (a)/Figure 4.8 and the (a)/Figure 4.11).

After the filter-bank processing of the 64 sub-band/filter-bank, output signals of several sub-bands are shown in the: (b), (c), (d), (e), (f), (g) and (h)/Figure 4.6. For the 3rd sub-band, the 24 sub-signals outputs of the NMF technique are shown in the 2nd to the 2(e)s/Figure 4.7. By applying of the exist prepared speaker clustering (the open-source “[GitHub](#)”) on these 24 sub-signals, there 4 speech signals. Two of them belong to the female, one by using the binary mask (the (c)/Figure 4.8) and the another belongs to the soft mask (the (d)/Figure 4.8). The other two speech signals belong to the male, one by using the binary mask (the (c)/Figure 4.11) and the another belongs to the soft mask (the (d)/Figure 4.11).

For the frequency domain comparison, spectrograms of the Figure 4.8 speech signals are analyzed in Figure 4.9 and

Figure 4.10. Spectrograms of the Figure 4.11 are analyzed in Figure 4.12 and

Figure 4.13.

Graphically, to compare between the binary sharp masking and the soft masking, the discontinuities of the output waveforms of the binary is obvious. In contrast, the soft masking produces little discontinuities on their waveforms for the output speech. Generally, these discontinuities and contentiousness speech appear in the (c)/Figure 4.8 and the (c)/Figure 4.11 or the binary and in the (d)/Figure 4.8 and the (d)/Figure 4.11 for the soft masks. Obviously, the Figure 4.19 shows these differences for one second (from 4 s to 5 s).

For the speech separation, the reference speech is called the targeted speech. There are several speech separation objective tests [25, 28]. The most reliable tests are the following four tests:

- The energy Source to Distortion Ratio SDR test is (by the decibels dB) is:

$$SDR (dB) = 10 \log_{10} \frac{\|S_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \quad (4-12)$$

where, S_{target} is the reference signal, e_{interf} is the interference error, e_{noise} is the noise error and e_{artif} is the artifact error. The error is the absolute distance between the output signal and the reference targeted signal.

- The energy Source to Interferences Ratio SIR test (by the decibels) is:

$$SIR (dB) = 10 \log_{10} \frac{\|S_{target}\|^2}{\|e_{interf}\|^2} \quad (4-13)$$

- The energy Source to Artifacts Ratio SAR test (by the decibels) is:

$$SAR (dB) = 10 \log_{10} \frac{\|e_{interf} + e_{noise} + S_{target}\|^2}{\|e_{artif}\|^2} \quad (4-14)$$

These ratios are measured by the decibels (dB). They have been defined and formulated in Chapter 1/1.10 **Subjective Test versus Objective Test** [25, 28].

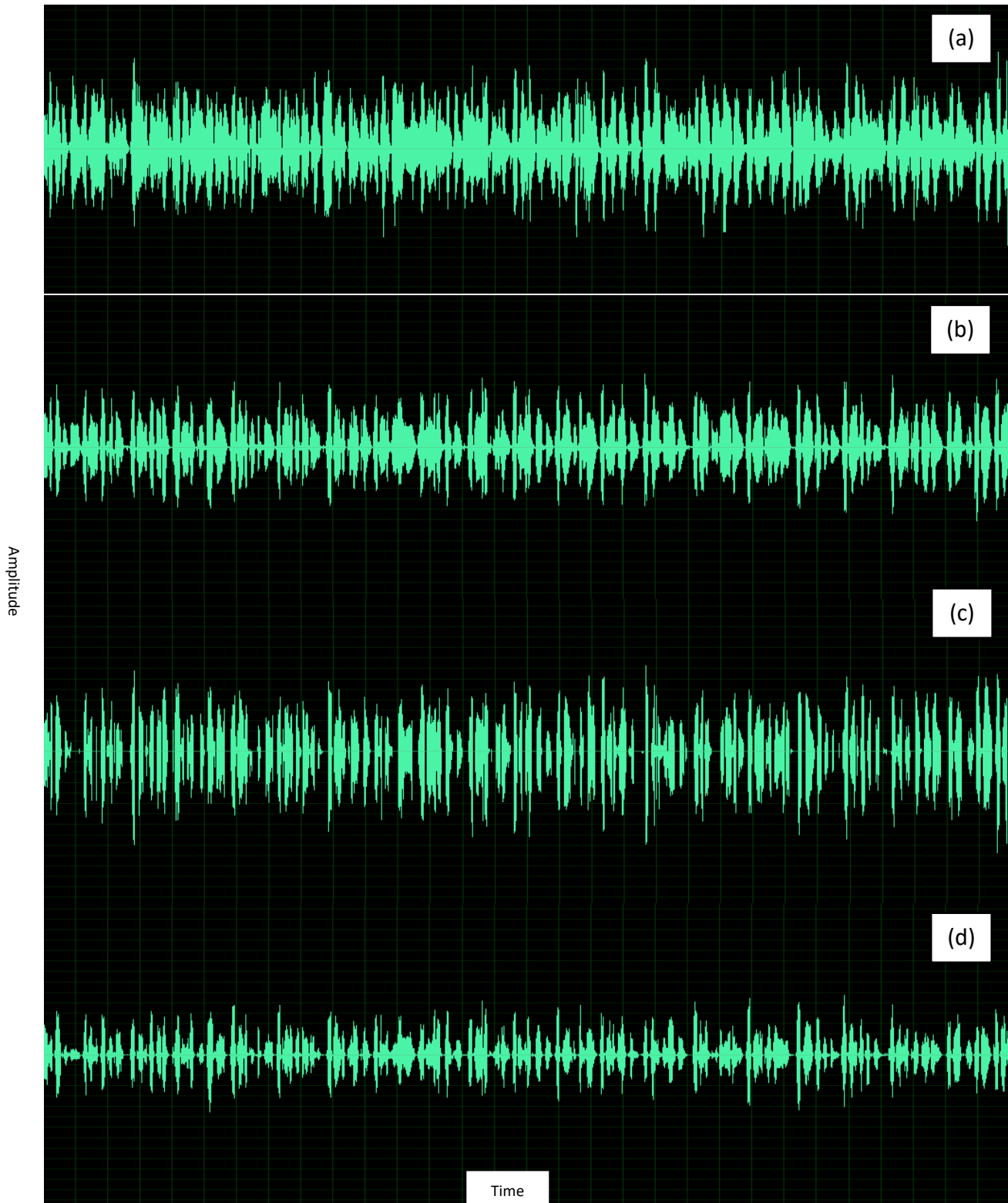


Figure 4.8 Speech signal waveforms of the Female (TIMIT). Speech signal waveforms of: the mixture speech FM (a), the female targeted-speech F (b), the separated female speech by the using of binary mask (c) and soft mask (d). There are horizontal-axes time-domain relationships between all the sketches.

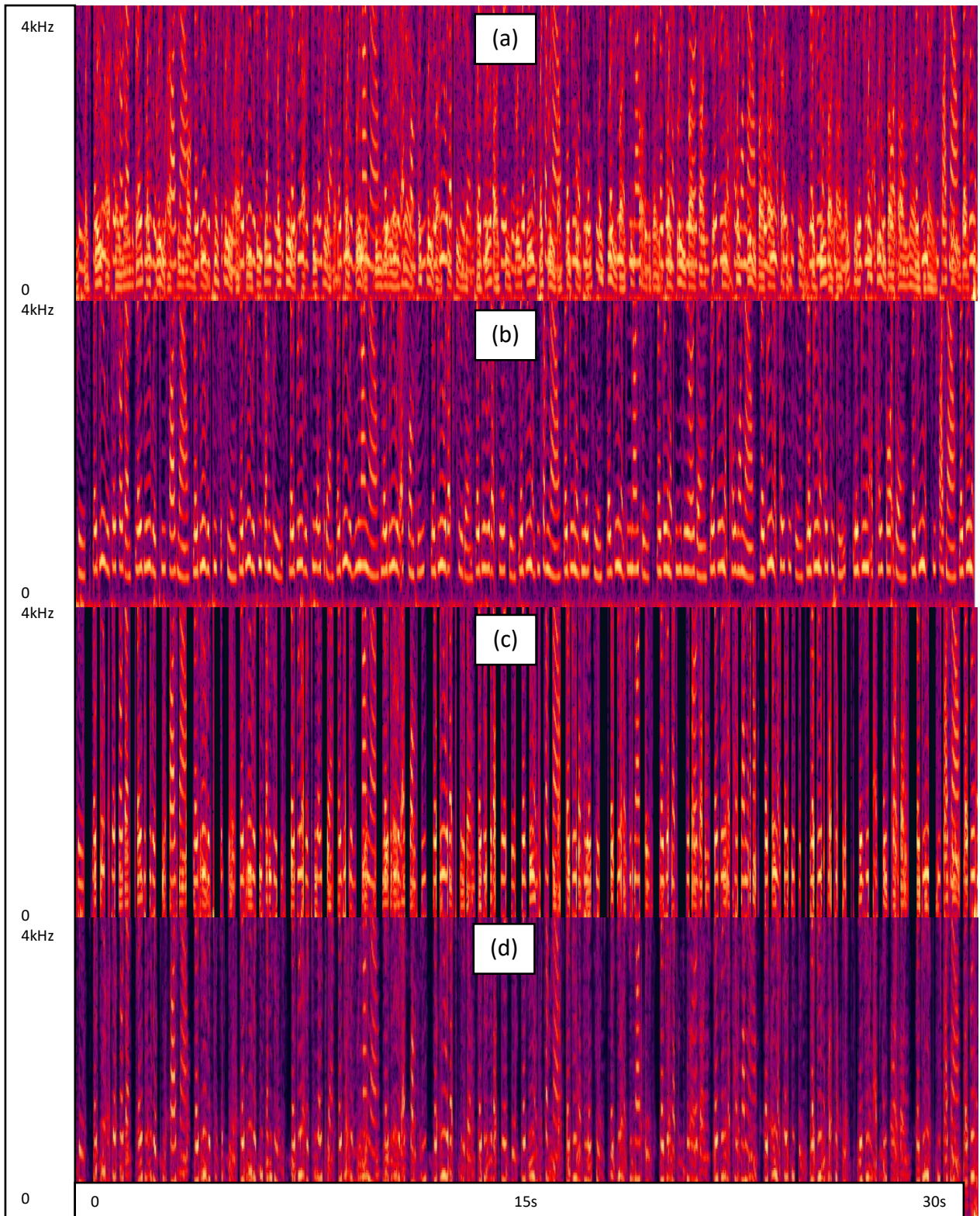


Figure 4.9 Spectrograms of the Female (TIMIT), for the 1 kHz range. Spectrogram of: the mixture speech FM (a), the female targeted-speech F (b), the separated female speech by the using of binary mask (c) and soft mask (d). There are horizontal-axes time-domain relationships between all the sketches.

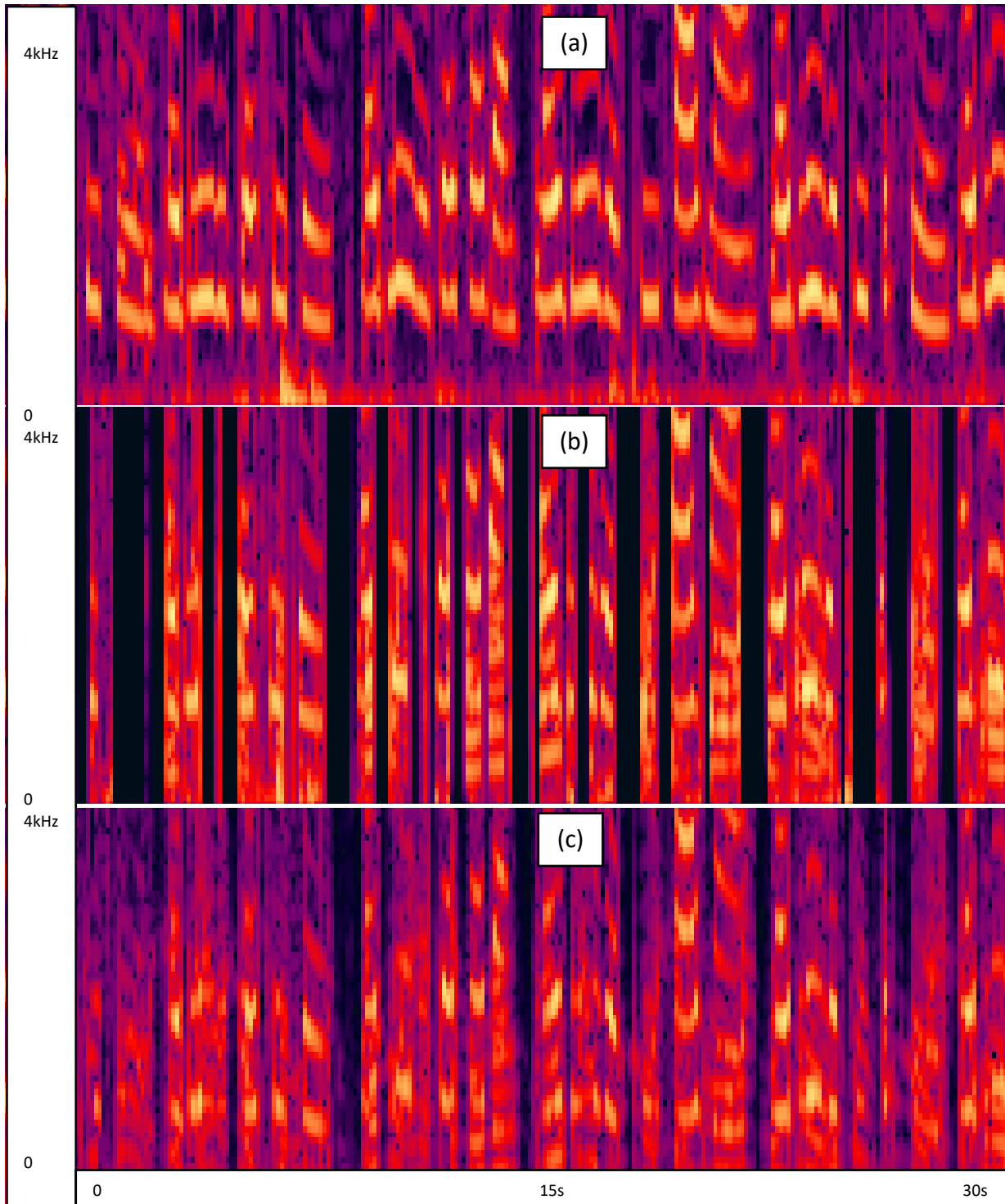


Figure 4.10 7s-500Hz T-F Spectrogram of the tested Female. The (a) is the targeted-speech of the F-TIMIT speech. The (b) is the recovered using the binary mask. The (c) is the recovered using the soft mask. There are horizontal-axes time-domain relationships between all the sketches.

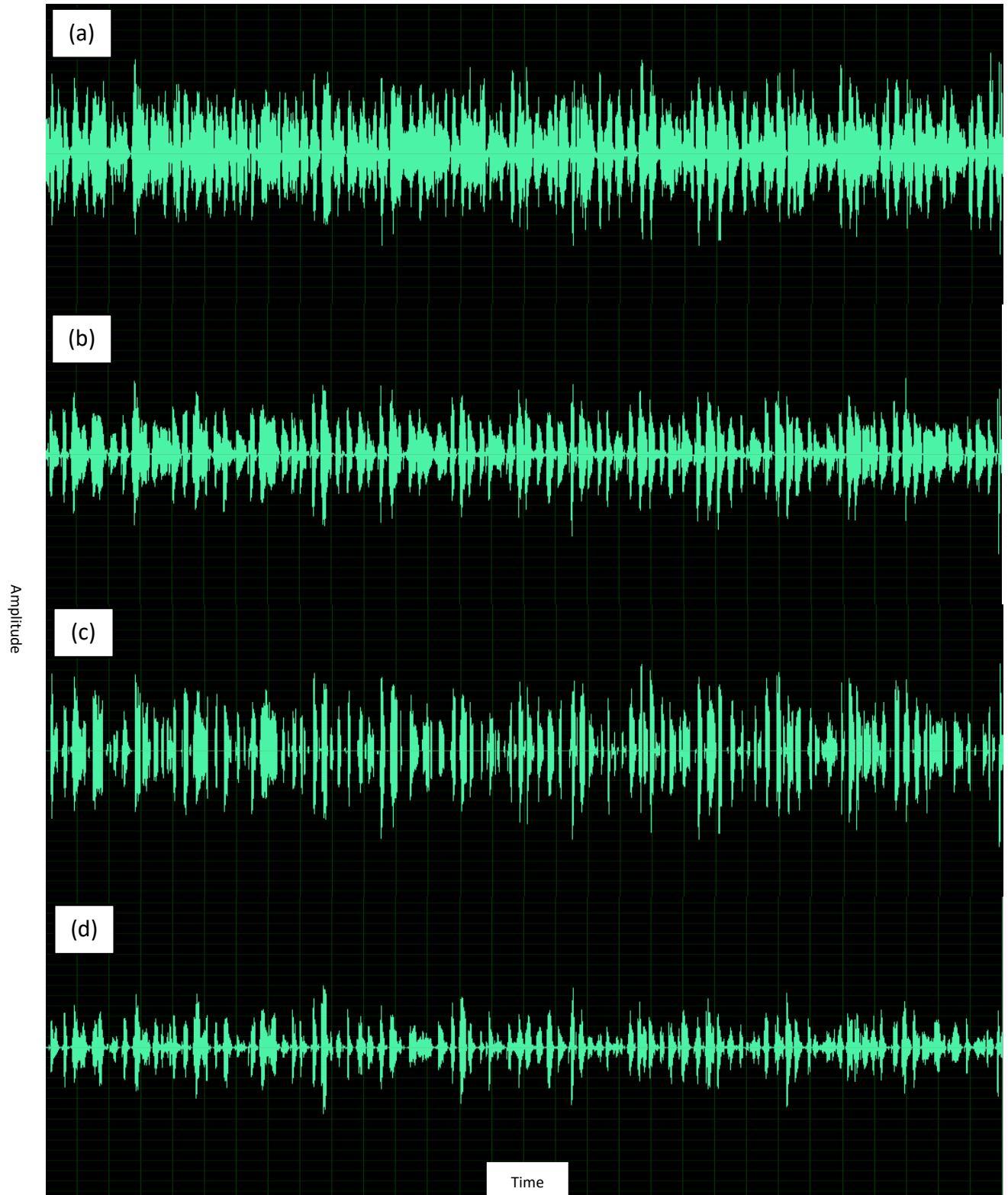


Figure 4.11 Speech signal waveforms of the Male (TIMIT). Speech signal waveforms of: the mixture speech FM (a), the male targeted-speech M (b), the separated female speech by the using of binary mask (c) and soft mask (d). There are horizontal-axes time-domain relationships between all the sketches.

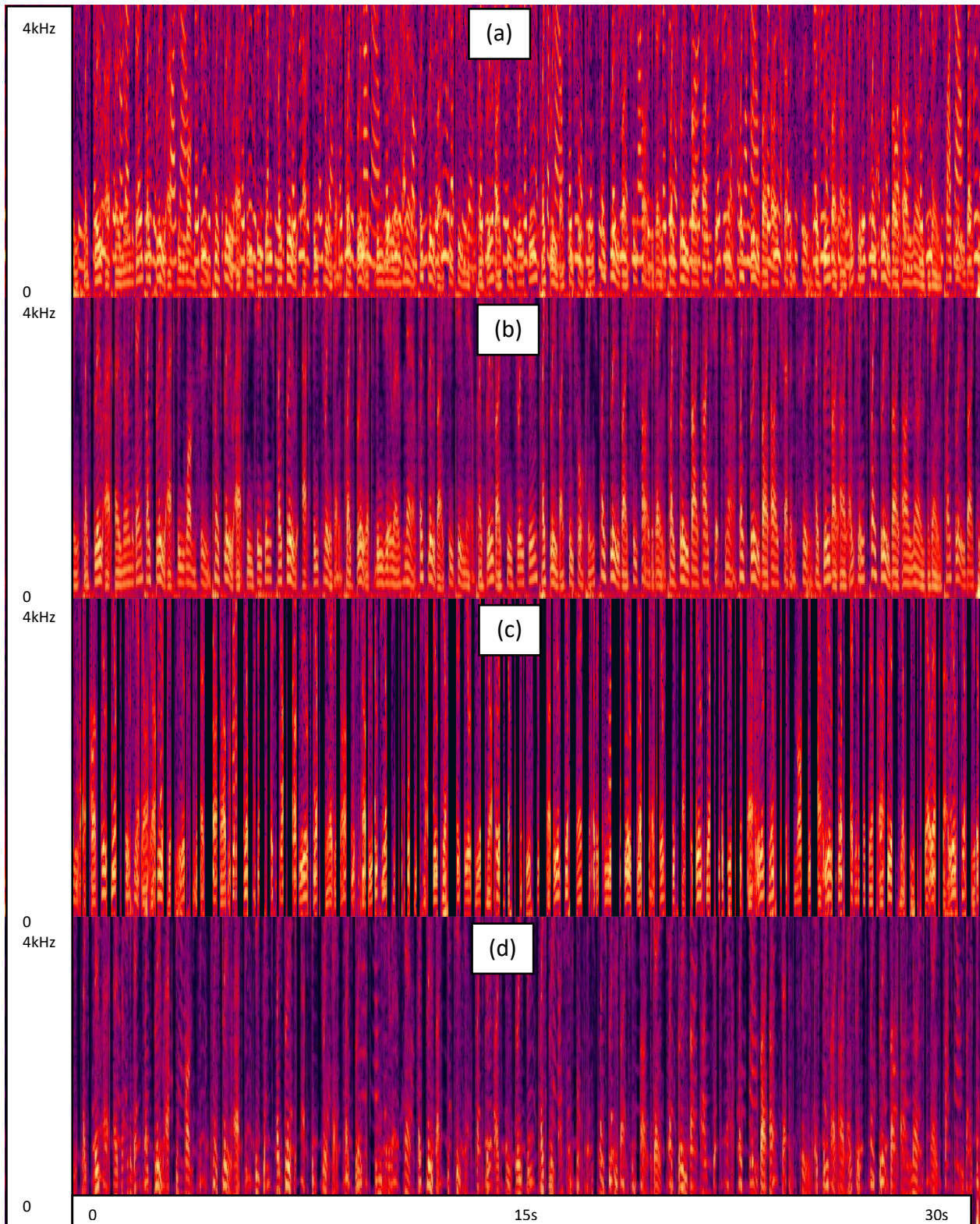


Figure 4.12 Spectrograms of the Male (TIMIT), for the 1kHz range. Spectrogram of: the mixture speech FM (a), the female targeted-speech M (b), the separated female speech by the using of binary mask (c) and soft mask (d). There are horizontal-axes time-domain relationships between all the sketches.

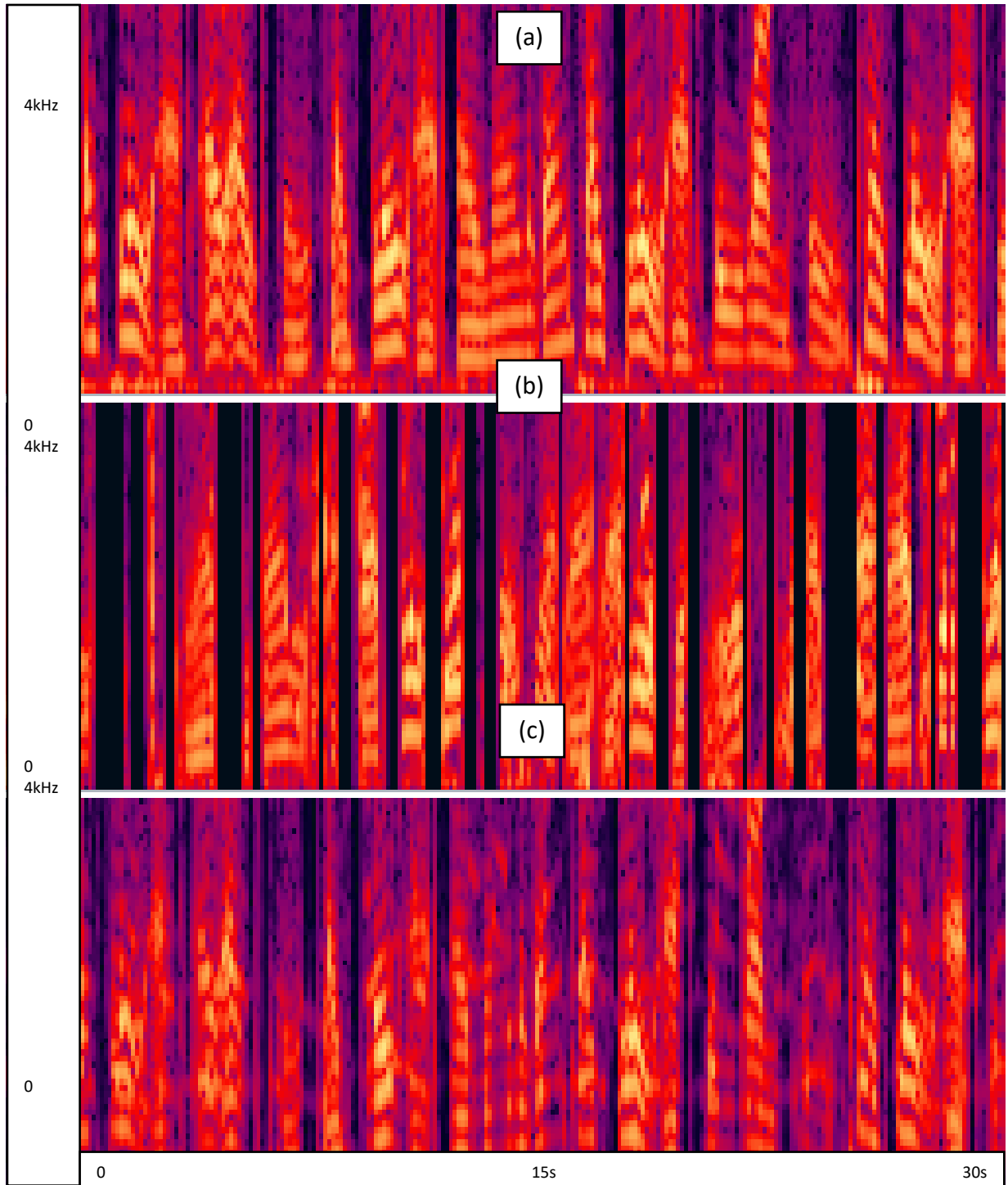


Figure 4.13 7s-500Hz T-F Spectrogram of the tested Male. The (a) is the targeted-speech of the M-TIMIT speech. The (b) is the recovered using the binary mask. The (c) is the recovered using the soft mask. There are horizontal-axes time-domain relationships between all the sketches.

The minimum values of the SAR, the SDR and the SIR are listed in the Table 4.1. The table is drawn, horizontally and vertically, in Figure 4.14.

Table 4.1 Minimum average values of the SAR, the SDR and the SIR ratios in dB for the output Female (F) and Male (M). The conversations are 17 Female-Male (FM), 17 Female-Female (FF), 17 Male-Male (MM) and the for all the 51 conversations (All); e.g. SAR-FB is the SAR for the Female output speech signal using the Binary mask.

	SAR-FB	SAR-FS	SAR-MB	SAR-MS	SDR-FB	SDR-FS	SDR-MB	SDR-MS	SIR-FB	SIR-FS	SIR-MB	SIR-MS
FM	1.91	3.91	1.00	3.84	1.09	2.99	0.14	2.14	7.33	7.58	9.09	8.46
FF	1.53	4.00	1.69	4.23	0.78	2.66	1.16	3.00	8.57	9.05	8.46	9.02
MM	1.60	3.92	1.38	3.63	0.53	2.24	0.88	2.54	8.03	8.14	8.30	8.20
All	1.68	3.94	1.36	3.90	0.80	2.63	0.73	2.56	7.98	8.26	8.62	8.56

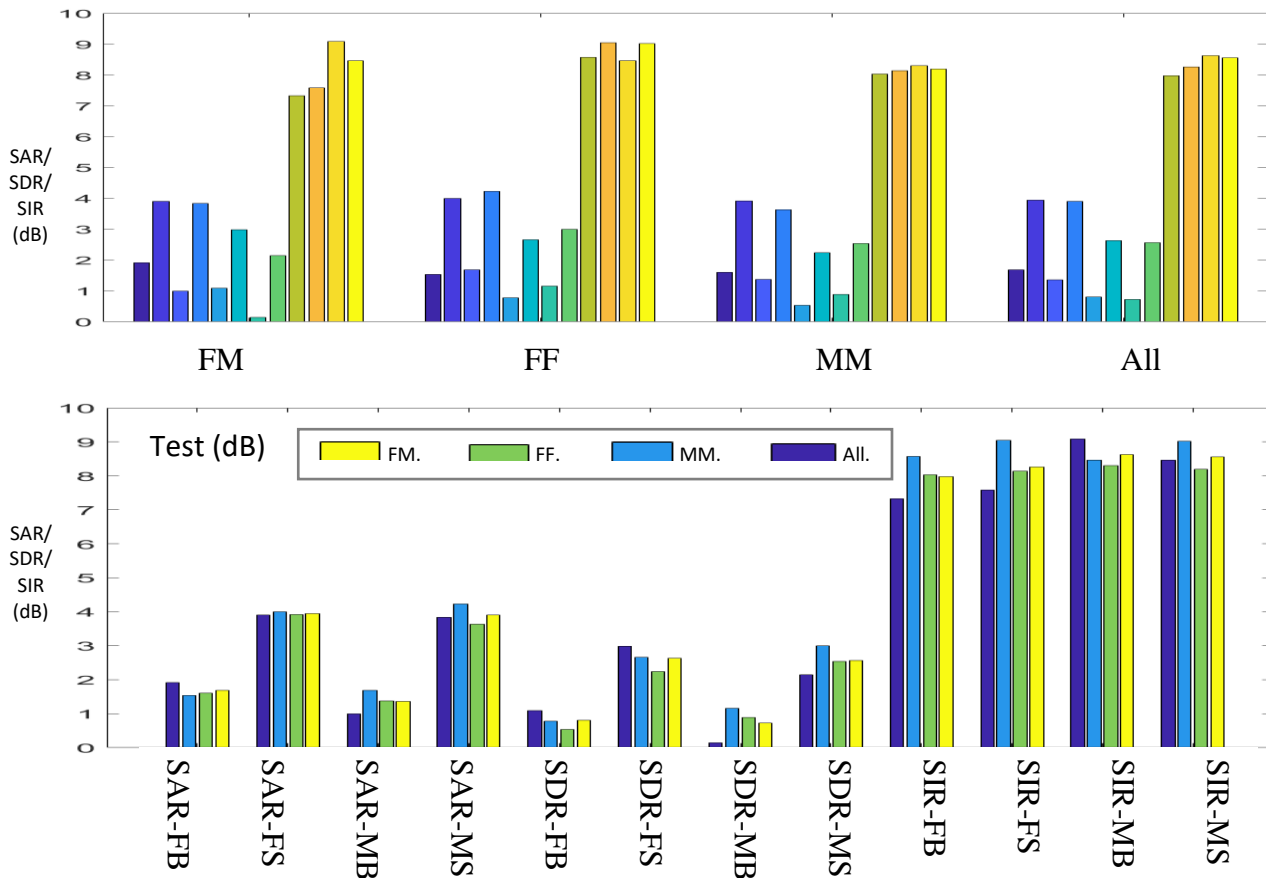


Figure 4.14 Graphic bars represent the minimum values of the objective tests. The above contains the FM, the FF, the MM and the All conversations collections. Each collection, sequentially from the left to right, has the following tests: SAR-FB, SAR-FS, SAR-MB, SAR-MS, SDR-FB, SDR-FS, SDR-MB, SDR-MS, SIR-FB, SIR-FS, SIR-MB and SIR-MS.

The maximum values of the SAR, the SDR and the SIR are listed in the Table 4.2. The table is drawn, horizontally and vertically, in Figure 4.15.

Table 4.2 Maximum average values of the SAR, the SDR and the SIR ratios in dB for the output Female (F) and Male (M). The conversations are 17 Female-Male (FM), 17 Female-Female (FF), 17 Male-Male (MM) and the for all the 51 conversations (All); e.g. SAR-FB is the SAR for the Female output speech signal using the Binary mask.

	SAR-FB	SAR-FS	SAR-MB	SAR-MS	SDR-FB	SDR-FS	SDR-MB	SDR-MS	SIR-FB	SIR-FS	SIR-MB	SIR-MS
FM	4.60	5.92	5.24	6.42	3.37	4.55	3.92	4.84	18.64	14.06	15.95	12.56
FF	5.12	6.34	6.79	5.95	4.27	5.24	4.59	5.60	19.19	15.87	17.08	14.69
MM	4.58	5.89	4.54	5.99	3.26	4.15	3.35	4.50	15.80	13.31	15.41	12.73
All	4.77	6.05	5.52	6.12	3.63	4.65	3.95	4.98	17.88	14.41	16.14	13.33

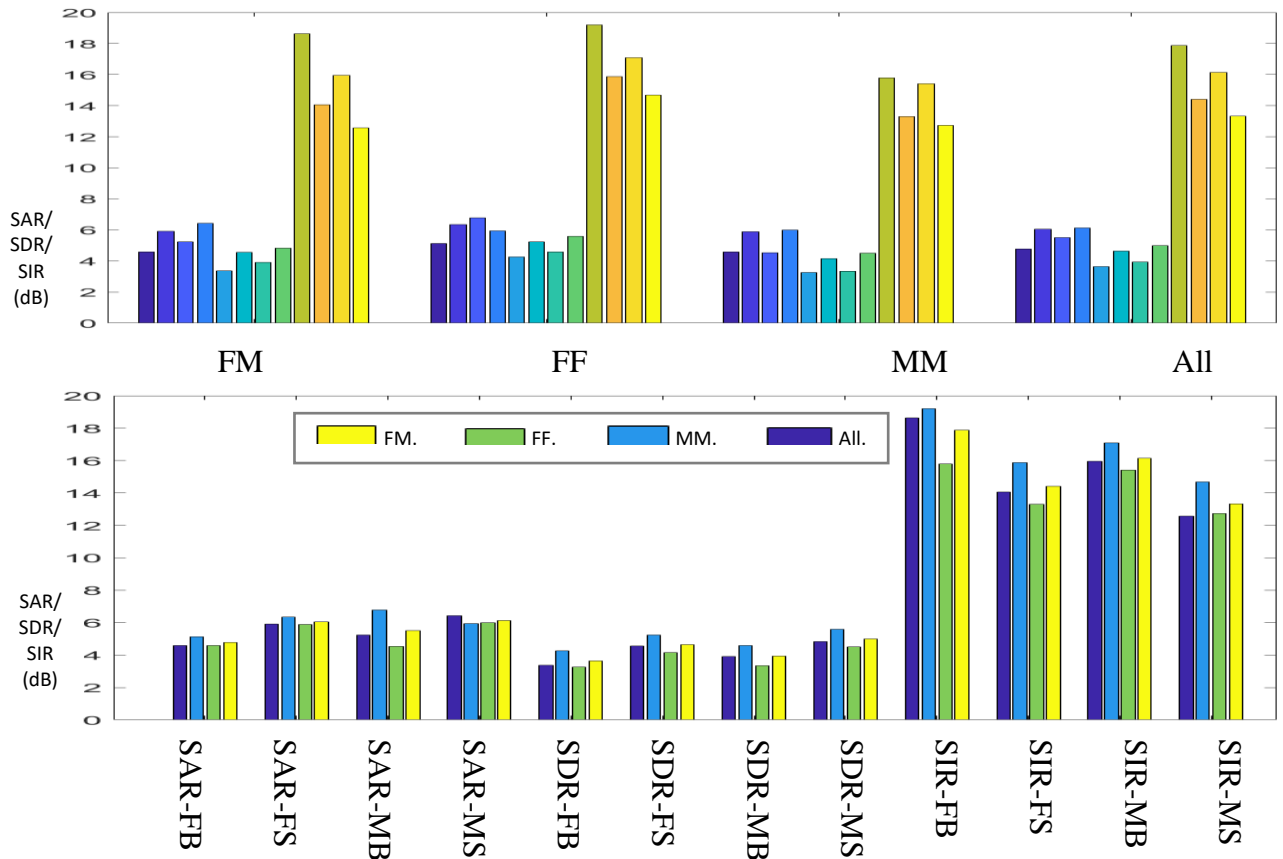


Figure 4.15 Graphic bars represent the maximum values of the objective tests. The above contains the FM, the FF, the MM and the All conversations collections. Each collection, sequentially from the left to right, has the following tests: SAR-FB, SAR-FS, SAR-MB, SAR-MS, SDR-FB, SDR-FS, SDR-MB, SDR-MS, SIR-FB, SIR-FS, SIR-MB and SIR-MS.

Neither the minimum nor the maximum values describe the truth behavior of the algorithm. The minimums and maximums appear the special good and bad cases. The average of them can appear the actual ability of the algorithm \pm the tolerance. The average values of the SAR, the SDR and the SIR are listed in the Table 4.3. The table is drawn, horizontally and vertically, in Figure 4.16.

Table 4.3 Average values of the SAR, the SDR and the SIR ratios in dB for the output Female (F) and Male (M). The conversations are 17 Female-Male (FM), 17 Female-Female (FF), 17 Male-Male (MM) and the for all the 51 conversations (All).

	SAR-FB	SAR-FS	SAR-MB	SAR-MS	SDR-FB	SDR-FS	SDR-MB	SDR-MS	SIR-FB	SIR-FS	SIR-MB	SIR-MS
FM	3.28	4.92	3.21	4.94	2.38	3.63	2.35	3.63	12.00	10.93	11.86	10.84
FF	3.79	5.20	3.65	5.09	2.91	4.01	2.84	4.02	12.65	11.64	13.11	11.77
MM	3.43	5.05	2.78	4.58	2.36	3.60	2.05	3.38	10.89	10.39	12.45	11.03
All	3.50	5.06	3.21	4.87	2.55	3.75	2.41	3.68	11.85	10.99	12.47	11.21

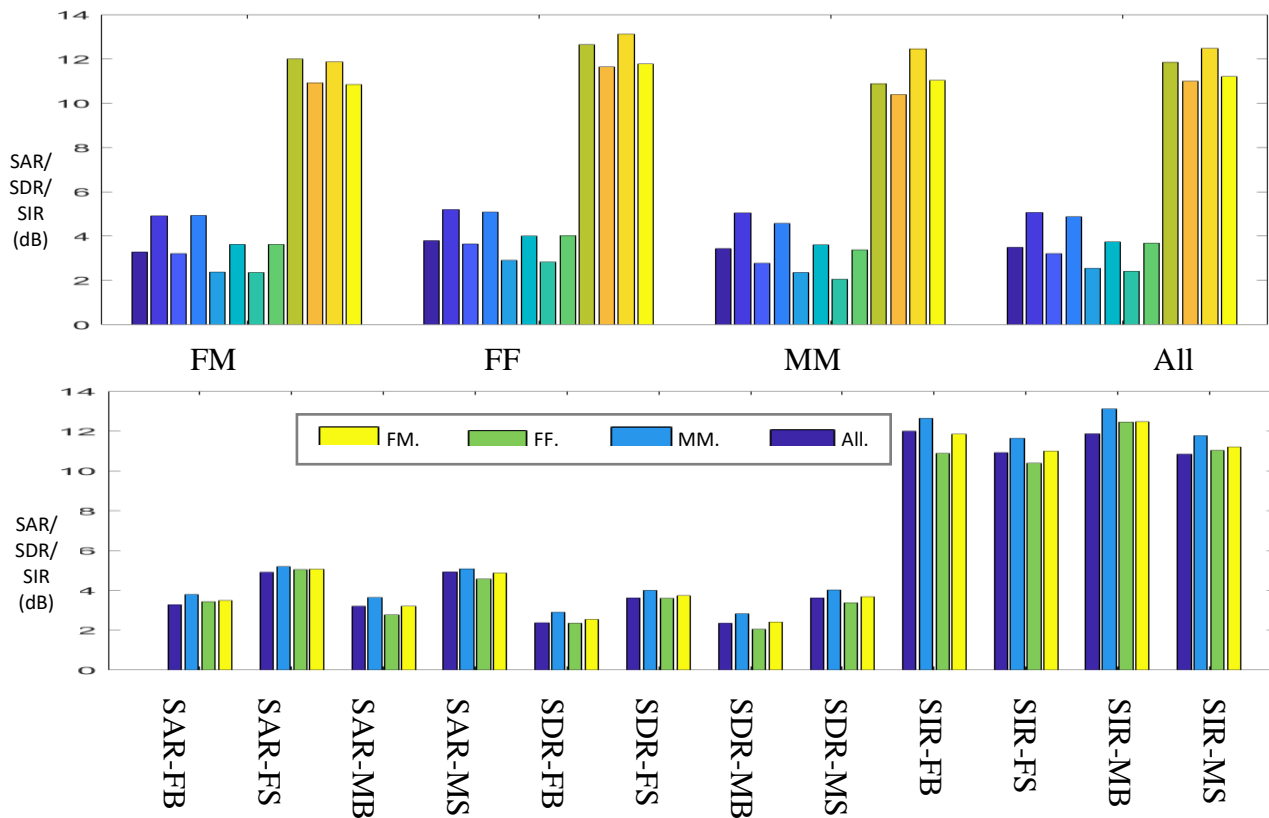


Figure 4.16 Graphic bars represent the average values of the objective tests. The above contains the FM, the FF, the MM and the All conversations collections. Each collection, sequentially from the left to right, has the following tests: SAR-FB, SAR-FS, SAR-MB, SAR-MS, SDR-FB, SDR-FS, SDR-MB, SDR-MS, SIR-FB, SIR-FS, SIR-MB and SIR-MS.

To detail the previous data and graph-bars, the Figure 4.16 can be redrawn in Figure 4.17. The figure has the collections of the conversations FM, FF, MM and All conversations.

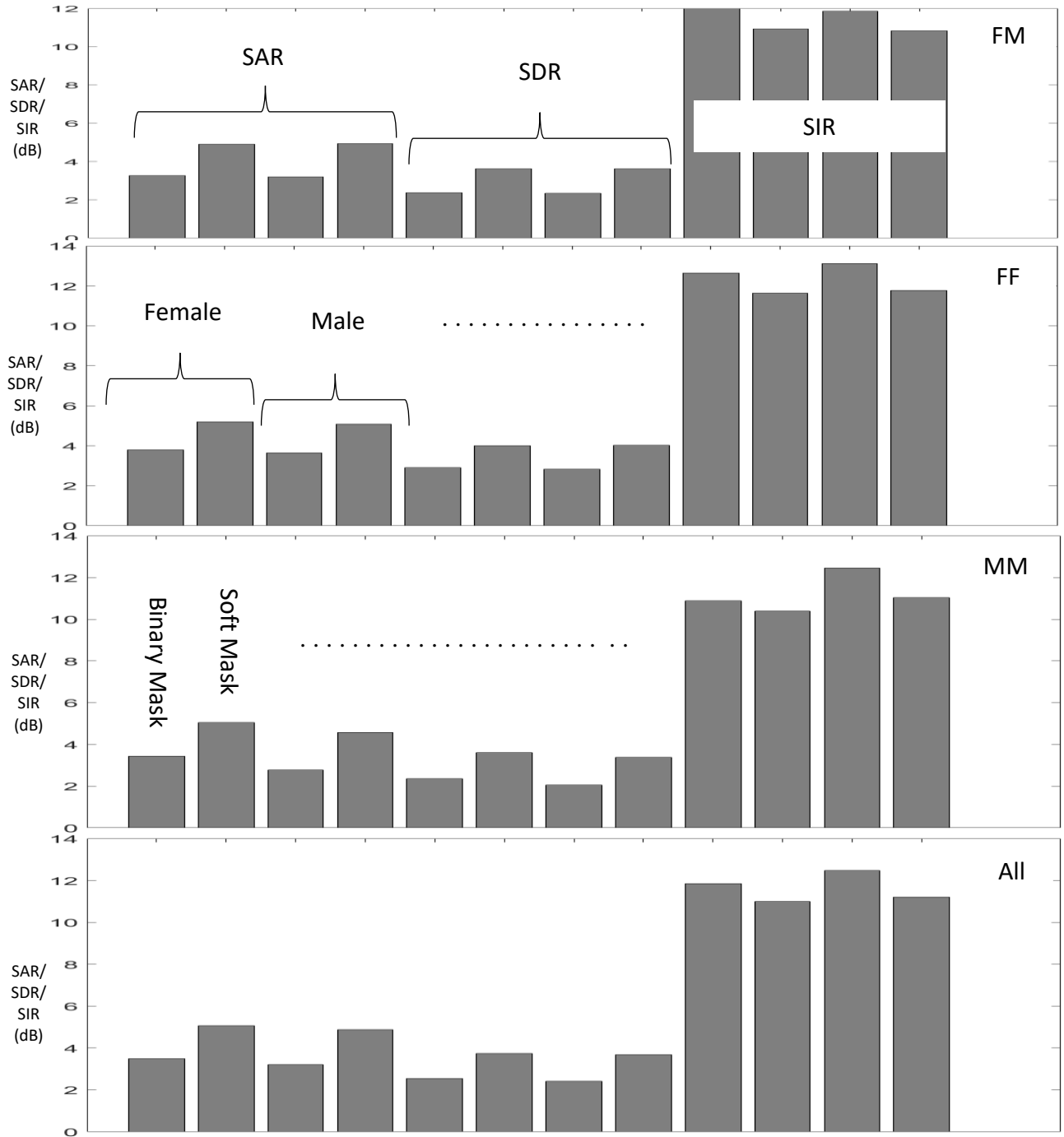


Figure 4.17 The average values of SAR-FB, SAR-FS, SAR-MB, SAR-MS, SDR-FB, SDR-FS, SDR-MB, SDR-MS, SIR-FB, SIR-FS, SIR-MB and SIR-MS (from left to right,). The (a) is for the FM conversations. The (b) is for the FF conversations. The (c) is for the MM conversations and the (d) is for the All conversations. There are horizontal-axes relationships between all the sketches.

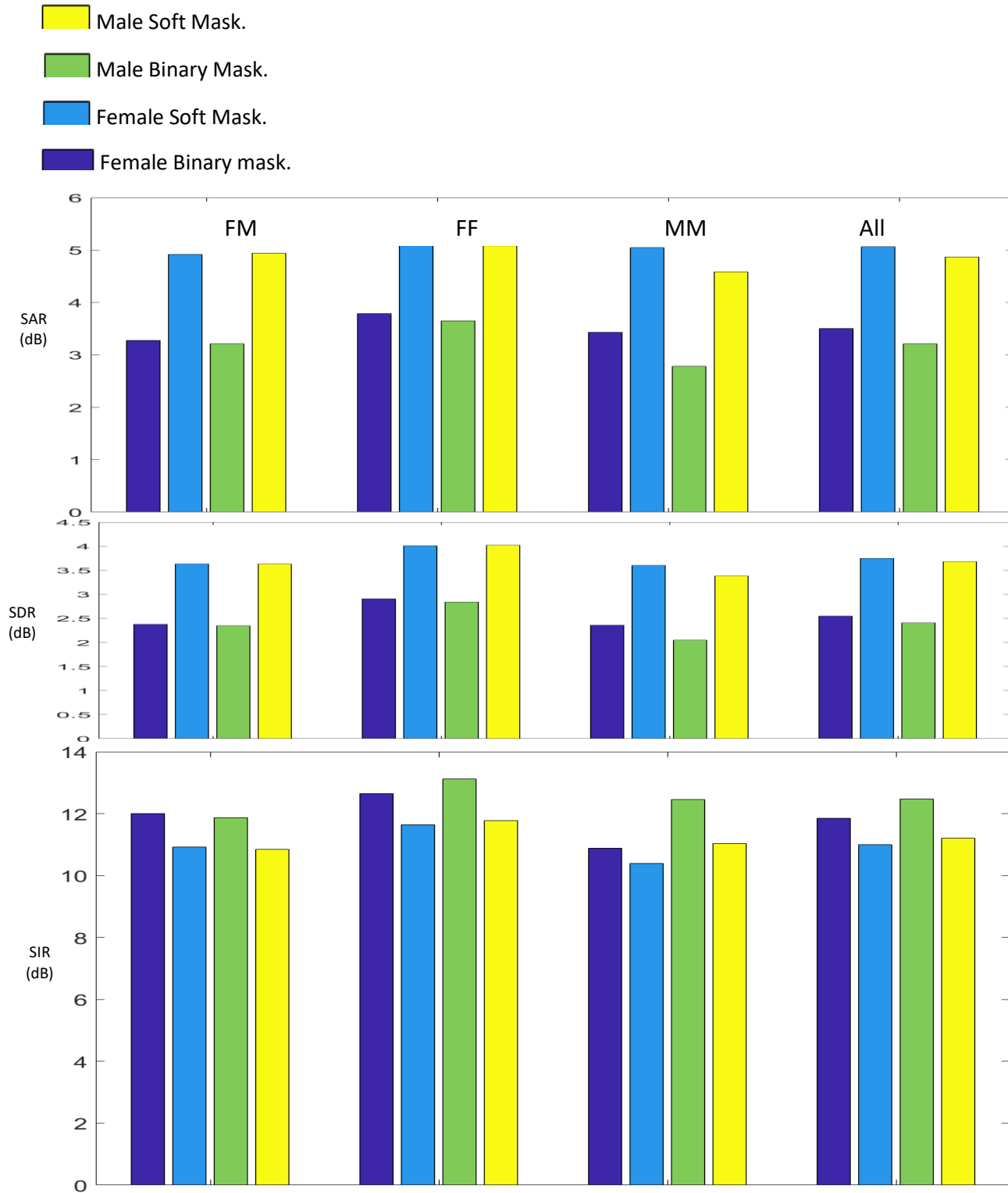


Figure 4.18 Graphic bars represent the average values of the SAR, the SDR and the SIR (dB). From left to right, the groups are: FM, FF, MM and All conversations. There are horizontal-axes relationships between all the sketches.

The average SAR, SDR and SIR tests of the 51 conversations are bar-graphed in the above Figure 4.18. The binary and the soft masks are listed. The bars represent the Female and Male output separated speech.

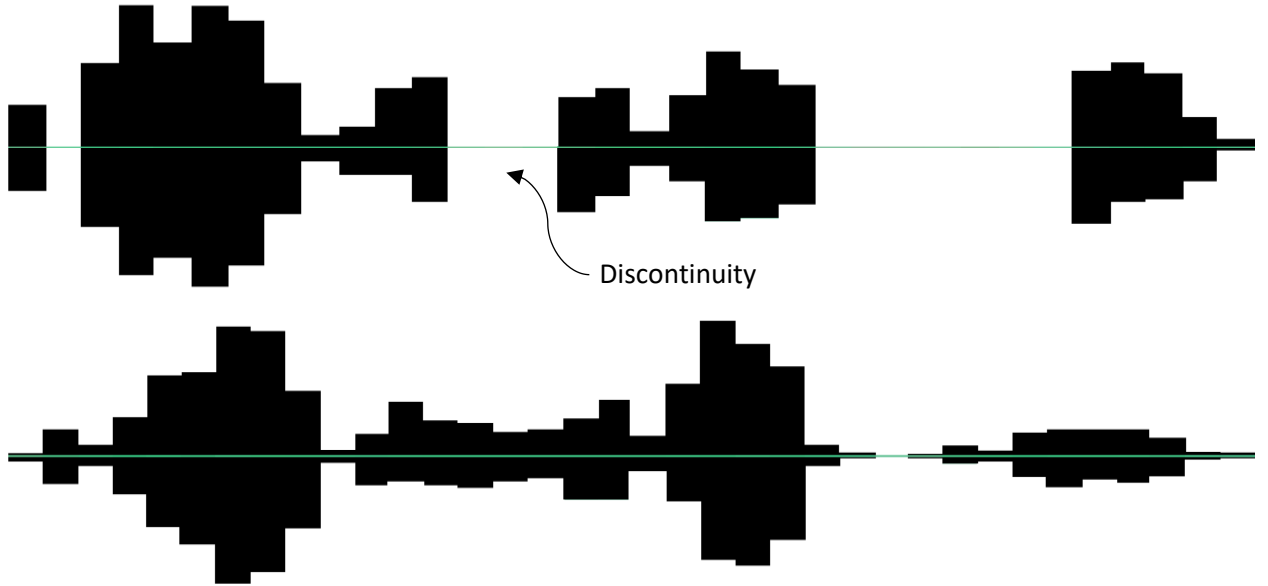


Figure 4.19 One-second waveforms to compare between the binary masking and the soft masking. The binary mask produces discontinuously periods of the waveform (the upper waveform), the soft mask does not (the lower). There are horizontal-axes time-domain relationships between all the sketches.

4.6 Comparison

The objective tests of this paper algorithm could be compared with the following current articles: For NMF-based speech separation, SAR is 4.5 to 8.5 dB, SDR is 1.0 to 4.0 dB [84]. For audio separation, Bayesian extensions to NMF are used: SDR is 0.1 to 3.9 dB for male, -0.7 to 8.6 dB for male; SAR is 1.4 to 5.2 dB for male, -0.9 to 3.4 dB for male; and SIR is 4.0 to 8.0 dB [85]. For cluster frequency basis functions of monaural sound source separation; the original source signals were used as a reference for the performance evaluation: clustering SDR, SIR and SAR are: Ckm is 0.80 dB, 10.96 dB and 3.30 dB; Cnmf is 2.89 dB, 12.72 dB and 4.59 dB; SNMFmap is 5.40 dB, 15.27 dB and 6.90 dB; SNMFmask is 8.94 dB, 23.69 dB and 9.72 dB [91]. SDR is 0.51 dB while the log-frequency spectrogram SDR is 2.8 dB; higher performance is attained by the Cochleagram with an average SDR is 8 dB [92]. The approach is for adaptive regularization of 2-D NMF for mixtures music and speech, SDR is 3.5. and 2.2dB; SDR is 3.5 dB [93]. The method against the uniform constant sparsity method: the improvement per source in terms of the SDR is 1.38 dB,

SAR is 1.38 dB, and SIR is 1.9 dB [94]. The comparisons are only with NMF-based blind speech and audio separation. The comparison does not involve the informed speech and audio separation. The comparison of this algorithm with other approaches (e.g. PCI or ICA) and/or the informed separation are not right.

For this paper algorithm: the SAR is 5.06 dB, the SDR is 3.75 dB and the SIR is 2.47 dB. According to the range of these tests, this chapter algorithm has moderate efficiency. The subjective tests by several listeners confirm this evaluation [22, 87].

4.7 Summary

Only one output segregated signal from Chapter 3 algorithms is processed in this chapter. The signal is the mixture speech of 2-speaker spontaneous conversation. This chapter blind speech separation achieves novel algorithm which are trying to recover the original speech of each speaker alone. The achievement is good compared with audio and sound speech separation by the NMF technique. Already, the speech separation by the NMF is evaluated as bad technique. In this chapter algorithm, the analysing of the filter-bank plus the identification of the speaker clustering, improve the badness of the NMF significantly.

The subjective tests of the resulting speech for the 2 speakers, denote clearly the acceptability of the separated speech. Like audio separation, the objective tests prove the ability of the NMF for the speech separation if the NMF technique is supported by other method(s).

This chapter algorithm need enough large number of sub-bands filter-bank and number of NMF sub-signals. More sub-bands and sub-signals increases the resolution of the analysis of the frequency domain content of the processed signal. This requirement is the weak-point against the algorithm. The required time to implement the algorithm depends entirely on the multiplication of number sub-band by number of sub-signal.

Chapter 5. Informed Speech Separation by Semi-Supervised Non-negative Matrix Factorization

5.1 Introduction

The main DSP objectives of the overlapped-speech conversation are the full speech isolation of each speaker from other speakers (the research focused only on the two speakers case for each conversation). That speech involves specific dialogue segments and share of the speech components during mixture segments. To achieve that job entirely, the input signal must be processed by: the overlapped-speech detection, the speaker diarization and the speech separation. In Chapter 3, the overlapped-speech detection is done successfully. The speaker diarization is implemented by invoking an existing toolbox. The speech separation processes the mixture speech, either without using the diarization outputs (blind speech separation), or with using the diarization outputs (informed speech separation). Chapter 4 describes novel speech separation algorithm without the using of the diarization output speech segments. In this chapter, the novel algorithm is arranged using the assistance of diarization outputs (i.e. the process is informed speech separation). The outputs of the speaker diarization is a generated database for that semi-supervised machine learning system. The database should be trained to easy the speech separation. The training supposes that there are two mixture signals, the first is the mixture speech output of the overlapped-speech detection (real-mixture). This signal is the observation input of the separation process, and should be separated into two signals, each one belongs to its speaker (because there are two speakers for the conversation). The second mixture signal is generated by the algebraic sum of the two outputs of the speaker diarization. For that, the second is called virtual-mixture speech signal, and the two outputs of the speaker diarization are virtual targeted-speech. To increase the configuration possibilities, that summation is implemented in the time and the time-frequency (spectrogram) domains. The NMF factorization of spectrogram matrices of the real-mixture and the virtual-mixture, is used for the training process. The analogy/coherence of the real-mixture with the virtual-mixture spectrogram matrices, are the approach to separate the real-mixture speech. Configuration of the output separated speech spectrogram matrix is congruent with the virtual targeted-speech spectrogram matrix configuration. Explicitly, number of DFT points for the two mixtures of the two targeted-speech is equal. Phase-angle of the mixtures and the conjugate-mirror property of the DFT guarantee the real part only of the IDFT. Each one of the soft and the hard masking divides the mixture speech into two speech signals, each signal belongs to its speaker. The

performances of the masks are fluctuating from objective test to others tests, and from conversation to other conversations. Due to that reason, merging them then choosing the efficient one is the last step to achieve the informed separation task [23, 118].

5.2 Functional Block Diagrams and Waveforms

Continuing with the details of the Figure 1.6 of the Chapter 3 and the Chapter 4, this chapter process is the second block (the second choice) of that overall system (see Figure 1.6). The Figure 5.1 and the Figure 5.2 are the general functional block diagram of the chapter system and its typical waveforms, respectively. The (a)/Figure 5.2 is the input of the system which is a spontaneous conversation of F with M . The (b)/Figure 5.2 is the first output of the detection, which is the dialogue signal (input of the speaker diarization). The (c)/Figure 5.2 and the (d)/Figure 5.2 are the virtual targeted-speech of F_v and M_v , respectively. The (e)/Figure 5.2 is the virtual-mixture speech signal. The (f)/Figure 5.2 is the second output of the overlapped-speech detection, which is the real-mixture speech signal. That real-mixture is the observation signal which should be separated into the real speech of F and real speech of M . The (g)/Figure 5.2 and the (h)/Figure 5.2 are those outputs, respectively. Virtual targeted-speech of F and M and virtual-mixture, assist the separation process [23, 118]. The assistance is done by analogizing the virtual-speech signals with the real-speech signals. Since the virtual-speech signals assist the process, the process of this chapter is informed speech separation (sometime, it is called semi-blind speech separation [161]).

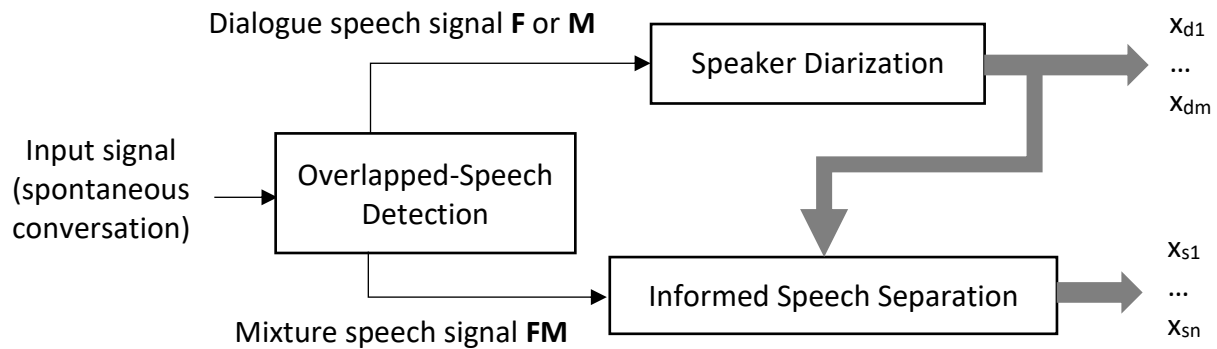


Figure 5.1 Chapter 5 overall system. The input is spontaneous conversation signal and the outputs are the individual speech signal of all the speakers (semi-supervised machine learning system).

This system is categorized as semi-supervised machine learning. The virtual-signals data could be named as database-like, because they are not already existed. They are generated from the overlapped-speech detection and the speaker diarization processes. These short-period signals have limited abilities compared with the long-period database. When these limited abilities are useful, the researchers were trying to exploit those useful capabilities.

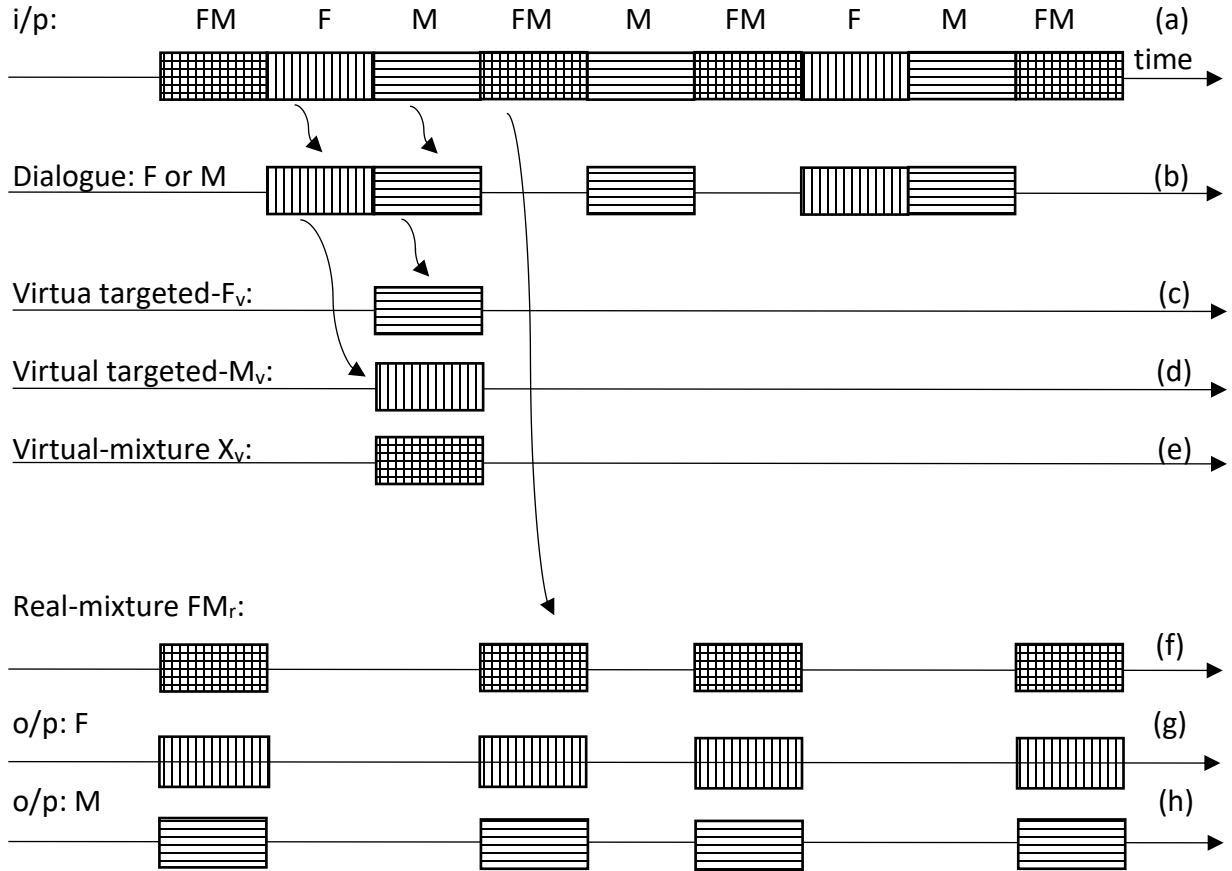


Figure 5.2 Chapter 5 arbitrary spontaneous conversation. The dialogue between Female F (vertically-lined) alone and Male M (horizontally-lined) alone. the mixture FM is both simultaneously (cross-lined). There are horizontal-axes time-domain relationships between all the sketches.

5.3 Informed Speech Separation Procedure

More details for the Figure 5.1 and the Figure 5.2 is illustrated in the Figure 5.3, where only the interaction between the diarization and the separation processes are focused. The dialogue output of the overlapped-speech detection has been processed by the diarization, which produces isolated speech segments of F and M. F segments are supposed as virtual targeted-speech of F (the

(c)/Figure 5.2), and M segments are supposed as virtual targeted-speech of M (the (d)/Figure 5.2). Algebraic sum of them produces virtual-mixture speech of them (the (e)/Figure 5.2). Input observation signal of the informed speech separation is the output mixture speech of the overlapped-speech detection. In this chapter, it has been named by the real-mixture speech signal. Training of the virtual-mixture and the virtual targeted-speech signal are done by using the conduction of the real with the virtual the NMF matrices spectrograms. To separate the real-mixture, two binary (hard) and non-binary (soft) masks are used. Switching from one mask to the other are done by the subjective-testers. To ensure if the two masks have different, useful or not useful masking, objective tests are used [23, 118].

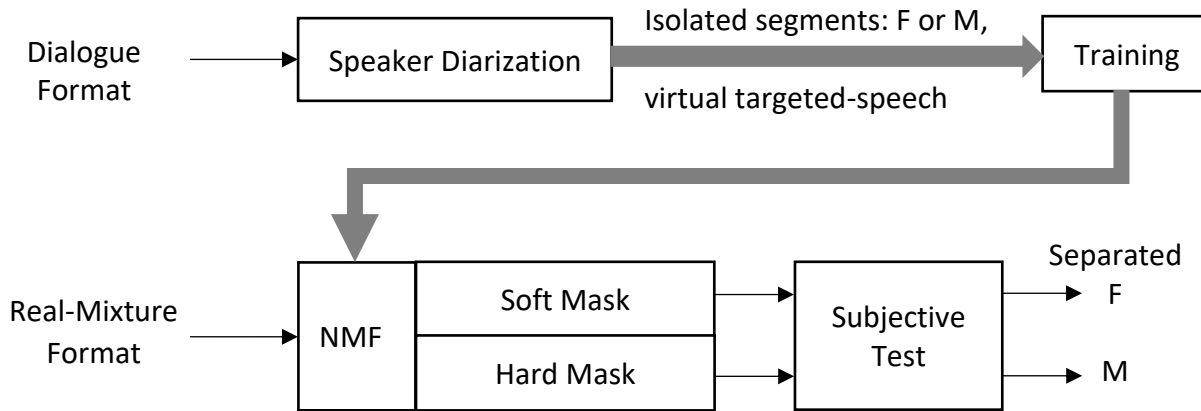


Figure 5.3 Functional block diagram of chapter 5 system.

5.3.1 Preparation of the Required Resources

During a spontaneous conversation, period of each individual dialogue segment and period of each mixture segment are random. During the spontaneous conversation, no one can determine the total time long of the individual segment, nor the total time long of the mixture segment. Almost, the period of any mixture speech segment is less than the summation of the periods for any individual dialogue speech. In this chapter, the worst probable case has been supposed, i.e. when the summation of periods for specific speaker equals one mixture segment. The ordinary spontaneous conversation guarantees that the total periods of any speaker, is longer than the longest mixture for that speaker with the other speaker.

For that, 30 second is chosen as the period of the real-mixture speech FM (i.e. relatively, very long period time for such format). This speech segment is important, because it is the processed

observation signal. According to that worst case, the long of each virtual targeted-speech segment (F and M) is 30 second. Since the virtual-mixture speech $f + m$ (F+M in frequency domain) is the weighted summation of the virtual targeted-speech segments, long of the virtual-mixture is 30 second also.

To prepare the suitable speech signals that meet that 30-s period, the real-mixture speech is the actual mixture output of the previous overlapped-speech detection. Already, the periods of each segment in Chapter 3 are 30 s. For the virtual speech signals, the dialogue output individual speech per segment is 30 s (Chapter 3). The speaker diarization process does not change those periods, so the virtual speech is 30 each segment [162]. For those reasons, input observation signal and semi-database of this chapter are already available. In this chapter, Chapter 3 overlapped-speech detection and the invoked speaker diarization, identically are achieved.

To avoid different sampling rates of the previous chapters, 8000 sample/s is used in this chapter speech segments. Since the resolution of the speech has no effect on the previous and the next processes, 16 bit/ sample is used to define the speech of each speaker. Like Chapter 3 and Chapter 4, the speakers are: F and M. For the first experiment, TIMIT female F and male M are checked [20, 163]. Other 33 speakers are experimented, later two-by-two (because each conversation has two speakers). The 33 speakers are arbitrary audio-book narrators. The books and the two speakers of each conversation are chosen arbitrary.

5.3.2 Training the Virtual Speech Signals

The first output signal of the overlapped-speech detection is the crude dialogue speech segments of the input spontaneous conversation (the (a)/Figure 5.2 is the input and the (b)/Figure 5.2). The invoked speaker diarization toolbox split those crude segments into collections of speech segments. The first collection belongs to the speaker F (the (c)/Figure 5.2). In time domain, it is symbolled by f . The second collection belongs to the speaker M (the (d)/Figure 5.2). In time domain, it is symbolled by m . These outputs of the speaker diarization, are the generated database which assist the main speech separation process. Homogeneously, that database should be inserted in and adapted with the input signal of the separation process (the (f)/Figure 5.2) [164].

In this chapter, virtual speech separation process is supposed, beside that original real speech separation process. To discriminate between them, the subscript (v) is suffixed the virtual speech signals symbols, and the subscript (r) is suffixed the real speech signals symbols.

The frequency-domain outputs of the speaker diarization are the virtual targeted-speech of F_v and M_v ; in time domain, they are f_v and m_v (the (c)/ and the (d)/Figure 5.2), i.e.:

$$[F_v] = FFT(f_v); \quad [M_v] = FFT(m_v) \quad (5.1)$$

where F_v and M_v are the time-frequency domain spectrograms of f_v and m_v . In time domain, the algebraic summation of those virtual targeted-speech of them is:

$$x_v = f_v + m_v; \quad (5.2)$$

where x_v is the time-domain virtual-mixture speech signal.

$$[X_v] = |FFT(x_v)| \quad (5.3)$$

where $[X_v]$ is the “time-frequency domain” spectrogram of the virtual-mixture speech signal (x_v).

$$[FM_v] = |[F_v]| + |[M_v]| \quad (5.4)$$

where FM_v is the time-frequency domain spectrogram of the virtual-mixture speech signal, by the adding of the time-frequency domain spectrograms of the virtual signals f_v and m_v . The above spectrogram matrices are prepared, in order to insert inside the NMF speech separation processing. For the virtual speech signals, both input virtual-mixture and the outputs virtual targeted-speech, are known:

$$[S] = [W] \times [H] \quad (5.5)$$

where $[S]$ is a spectrogram, $[W]$ is its Spectral-Basis matrix, and $[H]$ is its Activation-Weights matrix.

$$[FM_v] = [W_v] \times [X_v] \quad (5.6)$$

$$[W_v] = [FM_v] \times [X_v]^{-1}$$

The matrices $[FM_v]$ and $[X_v]$ are known. To find the matrix $[W_v]$, $[X_v]$ must be inverse-able. The dimensions of matrix must be square to find its inverse. Since there are relationships amongst the above matrices, the dimensions of all of them must be, explicitly checked [165].

The long of each f_v and m_v is 30 second. Since the sampling rate is 8000 sample/s, number of samples for each x_v , f_v and m_v is $N_t = 30 \times 8000 = 240,000$ samples. The speech frame is hopping

by 10 ms, so $T_h = 10 \text{ ms} = 0.010 \text{ s}$, therefore $N_h = 0.010 \times 8000 = 80$ samples each hop. Approximately, total number of the hopping frames $N_t = 240,000 / 80 = 3000$ frames. The width of the overlapping-windowed frame is $T_w = 32 \text{ ms} = 0.032 \text{ s}$, so $N_w = 0.032 \times 8000 = 256$ samples/frame. Number of DFT points equals number of these 256 samples. Since there are $((N_w/2) - 1)$ repeated mirror-conjugate sub-bands, number of sub-band is truncated to the rest actual sub-bands per frame: $N_w - ((N_w/2) - 1) = ((N_w/2) + 1) = 129$ sub-bands.

According to those calculations, each spectrogram matrix of $[FM_v]$, $[F_v]$, $[M_v]$, and $[X_v]$ has the dimension 129-by-3000. From the equation (5.6), $[W_v]$ has the 129-by-129 dimension, i.e. it is square matrix. To calculate the $[W_v]$ matrix elements, $[X_v]$ is not square matrix, its reciprocal cannot be calculated. For that reason, $[W_v]$ matrix cannot be calculated using the traditional matrix manipulations. To find the $129 \times 129 = 16,641$ elements of the $[W_v]$ matrix, 16641 equations must be set up. The alternative is the *pinv(.)* MATLAB function [166]:

$$[W_v] = [X_v] \times (\text{pinv}([FM_v])) \quad (5.7)$$

where *pinv(.)*, is the Moore-Penrose inverse (pseudo-inverse) of symbolic matrix, MATLAB function.

Mathematically, after (5.7) all the required matrices for the supervised-NMF process are available.

5.3.3 The Virtual Assists the Real Speech Signals

The second output of the overlapped-speech detection is the mixture speech segments of the main input spontaneous conversation. In order to recognize this mixture from the virtual mixture, this mixture is called real-mixture speech. In this Chapter, the mathematical terms of those real signals are suffixed by the subscript (*r*).

$$x_r = f_r + m_r \quad (5.8)$$

$$[S_r] = [W_r] \times [H_r] \quad (5.9)'$$

$$[X_r] = [W_r] \times [H_r] \quad (5.9)''$$

$$[FM_r] = [W_r] \times [H_r] \quad (5.9)$$

Since fm_r is the input known input signal with the 30-second speech segment. Its spectrogram matrix $[FM_r]$ has the dimension of 129-by-3000. The $[W_r]$ and $[H_r]$ are unknown-elements matrices. In this chapter, the matrix $[W_r]$ is approximated to the matrix $[W_v]$. There is no guarantee

that the approximation qualifies the equation (5.9) for the efficient solution which separate the $[FM_r]$ matrix into $[F_r]$ and $[M_r]$ matrices. The IDFT of the separated $[F_r]$ and $[M_r]$ matrices, are the separated f_r and m_r speech signals, which they are the real targeted-speech signals. The above configuration has been adopted after the failure checking of other several configurations. The know-how of that configuration is the kernel of the training process.

To solve the equation (5.9), the $[W_r]$ must be invertible. Since the matrix $[W_r]$ is square 129-by-129 elements, its reciprocal $[W_r]^{-1}$ can be calculated using the standard inverse MATLAB function (*inv(.)*) [166].

The phase-angle matrix of the transformed input signal $\text{DFT}(fm_r)$ is calculated and kept, then it is used to restore the phase-angle matrices of the separates speech signals. The restoration is done row-by-row for both F_r and M_r matrices. There is another choice to calculate the phase-angle matrix by the using of the phase-angle of the virtual mixture signal. For this Chapter that matrix is non-efficient alternative [167].

$$[W_r] = [W_v] \quad (5.10)$$

$$[H_r] = [W_r^{-1}] \times [FM_r] \quad (5.11)$$

$$[\varphi_{xr}] = \text{Phase_Angle}([X_r]) \quad (5.12)$$

To construct the separated targeted-speech signal, the 129 sub-signals $((N_w/2) + 1)$ of the NMF is separated one-by one into two signals. Phase angle is restored to each line using the invers step of the beginning step. Each separated signal in the frequency domain is complement by it mirror-conjugate complex values. The Separation decisions is done according to the measurements of the hard and/or the soft masks [23, 118].

5.3.4 Soft and Binary Masking

By the above NMF, 129 sub-signals are generated from the 30-second real-mixture segment. The long of each sub-signal is 30 second. The following is the sequence of that generation. The multiplication of each n column of the $[W_v(1:29, n)]$ matrix by the n row of the $[H_r(n, 1:129)]$ matrix, produce the n magnitude 129-by-129 spectrum. Each n from 1 to 129, produces 129-by-129 spectrogram matrix. The recovering of the phase-angle is done by the multiplication of each matrix with the phase-angle matrix $([\varphi_{xr}])$ [167]. To complete the full 256-point spectrum, their mirror conjugate must be added. The results are 129 matrices. Each matrix has 256-by-256 complex

value elements.

The masking techniques are the methods for making the separation decisions. The masking depends on the measured Euclidian-distances from these matrices to specific references. Since the system is semi-supervised ML, and there is useful information about the speakers. That information is used as the references. The references are the spectrogram-matrices of F and M, i.e. $[F_v]$ and $[M_v]$. The distances are the absolute values of 129-by-129 subtraction calculations. For each n (from 1 to 129) sub-signal, there are two 129-by-129 distances, the first distance for the F (F_{Ed}), and the second for the M (M_{Ed}).

In the masking, the share of F is inversely proportional with F_{Ed} , and the share of M is inversely proportional with M_{Ed} . There are two masking techniques to divide the sub-signal by that sharing. The first mask is the binary mask, which belongs all the element to the nearest speaker. The share of the farrest speaker is nulled (zero). The second mask is the soft mask, which belongs the largest part to the nearest speaker and belongs the smallest part to the farrest speaker. For the soft masking, the Euclidian distances are based on either the magnitude values or the power values. More details are detailed under the subtitle “**Chapter/1.11 Masking**” in the thesis [29, 30, 153].

According to the above description, the masking separates 129×129 of each sub-signal, i.e. each one of the $129 \times 129 \times 129$ elements is shared between F and M, using the two masking techniques. By the roughly statistical approximation, the binary hard masking produces 50% zero elements of both 129 F and 129 M matrices. The matrices which are produced by the soft masking do not have such merit [23, 118].

5.3.5 Exploiting both Masks

The previous masking step avails two separated speech signal for each speaker F and M. The objective testing denote that the quality of the resulting separated speech fluctuates from speaker to other speakers. For specific speaker, the performance fluctuates as well from conversation to other conversations. Both the binary hard masking and the soft masking fluctuate for different speakers and different conversations [168].

The checking involves the objective tests: the energy Source to Artifacts Ratio (SAR), the energy Source to Distortion Ratio (SDR) and the energy Source to Interferences Ratio (SIR) (see **1.10 Subjective Test versus Objective Test**). The fluctuation phenomenon involves those objective tests. For example, when the SAR test indicates positively for specific output speech, the other

SDR and/or the SIR indicates negatively. For specific test (e.g. SDR), the output separated F speech has different response compared with the output separated M speech. The test averages for 341 conversations denote that neither the soft mask alone nor the hard mask alone, satisfies the accepted results. Although the performance different is not good indication, but for this algorithm the different is exploited to improve the final output separated speech [29, 30, 153]. The exploitation is done by simple optimization step. Instead of using one mask, both the hard and the soft masks are used together. Figure 5.4 block diagram illustrates that optimization step.

To realize Figure 5.4 block diagram, the outputs of the masks must be objectively tested.

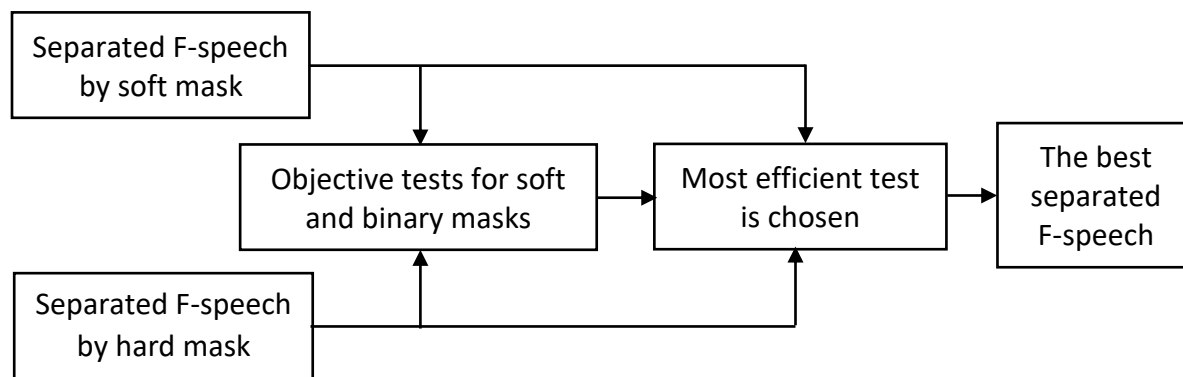


Figure 5.4 Optimization functional block diagram to improve the fluctuating of the objective tests.

According to Chapter 1/1.10 **Subjective Test versus Objective Test**, the targeted-speech must be available to calculate any one of those tests. The targeted-speech is the aim of the speech separation researchers. So, the block diagram is not realizable for the users. Although the block diagram is not applicable for the users, but it is very important for the researchers. If the diagram provides efficient output separated speech for the researchers, approximately it can be alternatively by the subjective test of any user. For example, the algorithm of this chapter passes this optimization block successfully for the researcher. That successful check enables the user to use such algorithm by the choosing the best output separated speech. That's mean, by his hearing the user selects the favorite separated speech, the soft masked or the hard masked.

By the exploiting of that optimization, the performance efficiency is improved significantly for 341 simulated conversations. More numerical details about that optimization step is plotted and tabulated in the result subtitle.

5.4 Experiments

The above detailed algorithm is simulated using MATLAB the DSP environment, AUDACITY the DSP speech editor and the other required software. The real and the virtual speech signals, already are available from the outputs of the overlapped-speech detection algorithm and the speaker diarization toolbox. Duration of each segment is 30 second. This time long is enough to express the true ability for the previous chapters and this chapter algorithm and algorithms.

The virtual speech f_v is algebraically summed to m_v to generate the virtual mixture speech signal x_v . Using FFT of f_v , m_v and x_v , the spectrogram matrices $[F_v]$, $[M_v]$ and $[X_v]$ are calculated. The summation of $[F_v]$ and $[M_v]$ matrices results the spectrogram matrix $[FM_v]$. Using the equation (5.7), the virtual Spectral-Basis matrix $[W_v]$ is calculated by the calling of the *pinv(.)* MATLAB built-in function [166].

For the real mixture, at first the real-speech Spectral-Basis matrix $[W_r]$ equals the virtual-speech Spectral-Basis matrix $[W_v]$. The equalization is approximately, and done by the previous supposing. Since $[W_r]$ is square matrix, its inverse is found easily. The Activation-Weights matrix $[H_r]$ of the real mixture is calculated using the equation (5.11). Phase angle matrix of the mixture speech should be prepared for the NMF factorization process.

The NMF, the masking and the separation, interacting processes are manipulated together. The NMF factorize the real mixture into 129 sub-signals. For each sub-signal, the Euclidian distance from its spectrogram to the references. The references are the virtual targeted-speech $[F_v]$ and $[M_v]$ spectrograms. These distances make the mask decisions. Both the hard binary and the soft masks shares the 129-by-129 points to F and M. The sharing is repeated for the 129 sub-signals.

The four-resulting output separated speech signals are tested subjectively and objectively. According to the objective tests, the outputs are optimized, i.e. choosing two output speech signals and neglect other two speech signals. Cross-checking between the objective optimization and the subjective optimization confirms the best separated speech signals.

Almost, the chosen parameters of the processing are similar those of Chapter 3 and Chapter 4. Since the sampling frequency is 8000 sample/s, the overlapping-window is 32 ms, i.e. 256 samples per frame. The hopping time is 10 ms, i.e. 80 samples each hop. The window function is the standard Hanning scaling window. The dimensions of the resulting spectrogram are 129-by-3000.

5.5 Results and Tests

The above simulation has been repeated 341 times for 341 spontaneous conversations. The conversations are prepared using 35 different speakers. Two of them are the Female and the Male of the standard TIMIT speech and audio library. Other 33 speakers are arbitrary audio-book well-known narrators. Figure 5.5 and Figure 5.7 show speech waveforms of: mixture, targeted, recovered by soft and recovered by binary mask, of F TIMIT and arbitrary M narrator, respectively. Figure 5.6 and

Figure 5.8 show speech spectrograms of: mixture, targeted, recovered by soft and recovered by binary mask, of F TIMIT and arbitrary M narrator, respectively.

The subjects of the conversations are different depending about the narrated book. The narrators are females and males. Total number of available conversations are $35 \times 36 \div 2 = 630$, but arbitrary 341 are enough to cover the requirements [23, 117]. For each conversation, there are 4 separated speech signals. Two of them belong to F and other two to M. One of each two is separated by the binary mask and the other by the soft mask. The user should choose the best separated speech signal, depending on his acceptance for the best one of the two output speech. This step is done by the user for both speakers the F and the M. To ensure that there is a clear difference between the two masks, objective tests can check them. For the first speaker, the objective tests are:

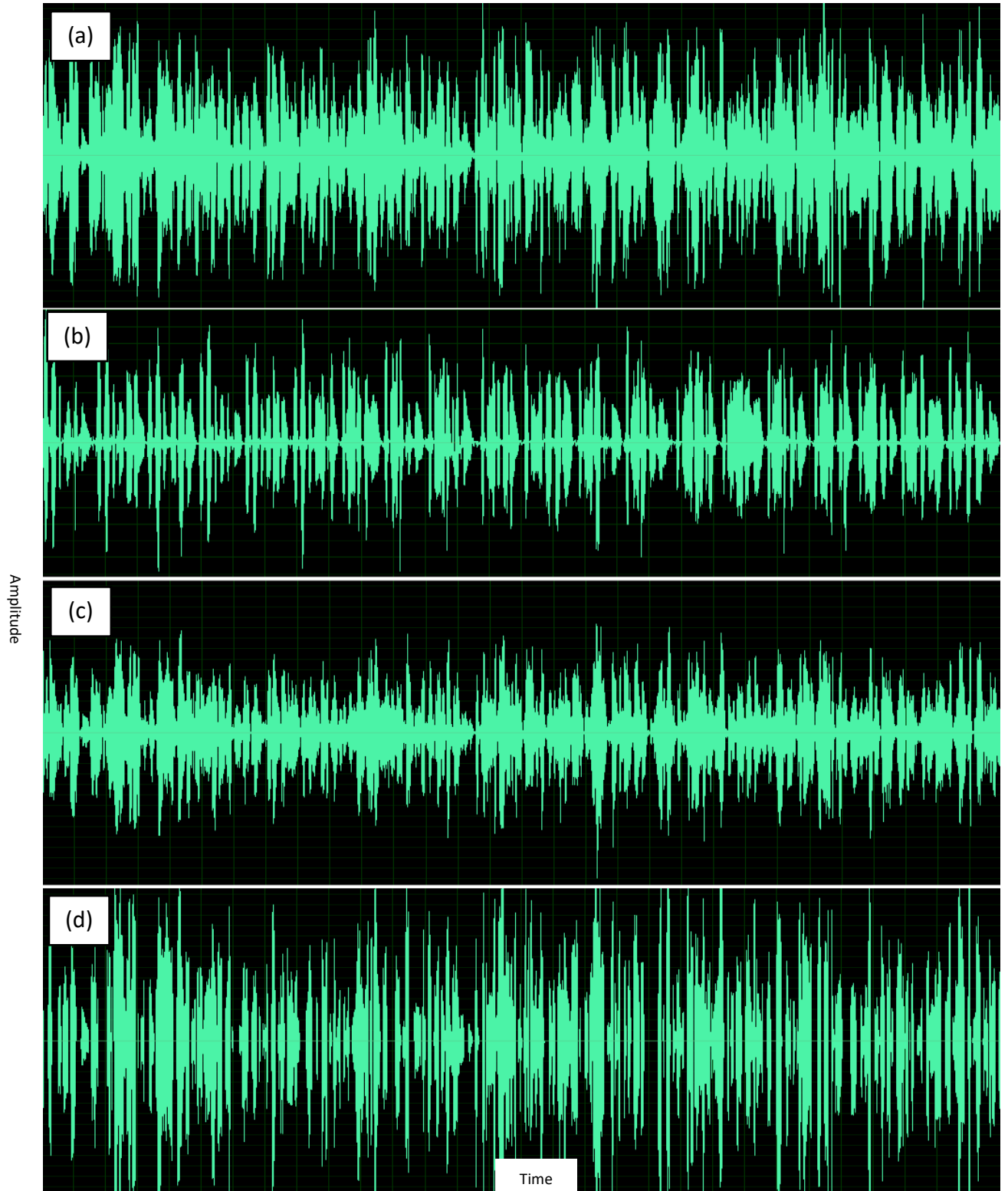


Figure 5.5 Waveforms of mixture, targeted and recovered speech of the 1st speaker. She is the F of TIMIT. The (a) is the mixture speech. The (b) is the targeted-speech. The (c) is the recovered-speech using the soft mask. The (d) is the recovered-speech using the binary mask. There are horizontal-axes time-domain relationships between all the sketches.

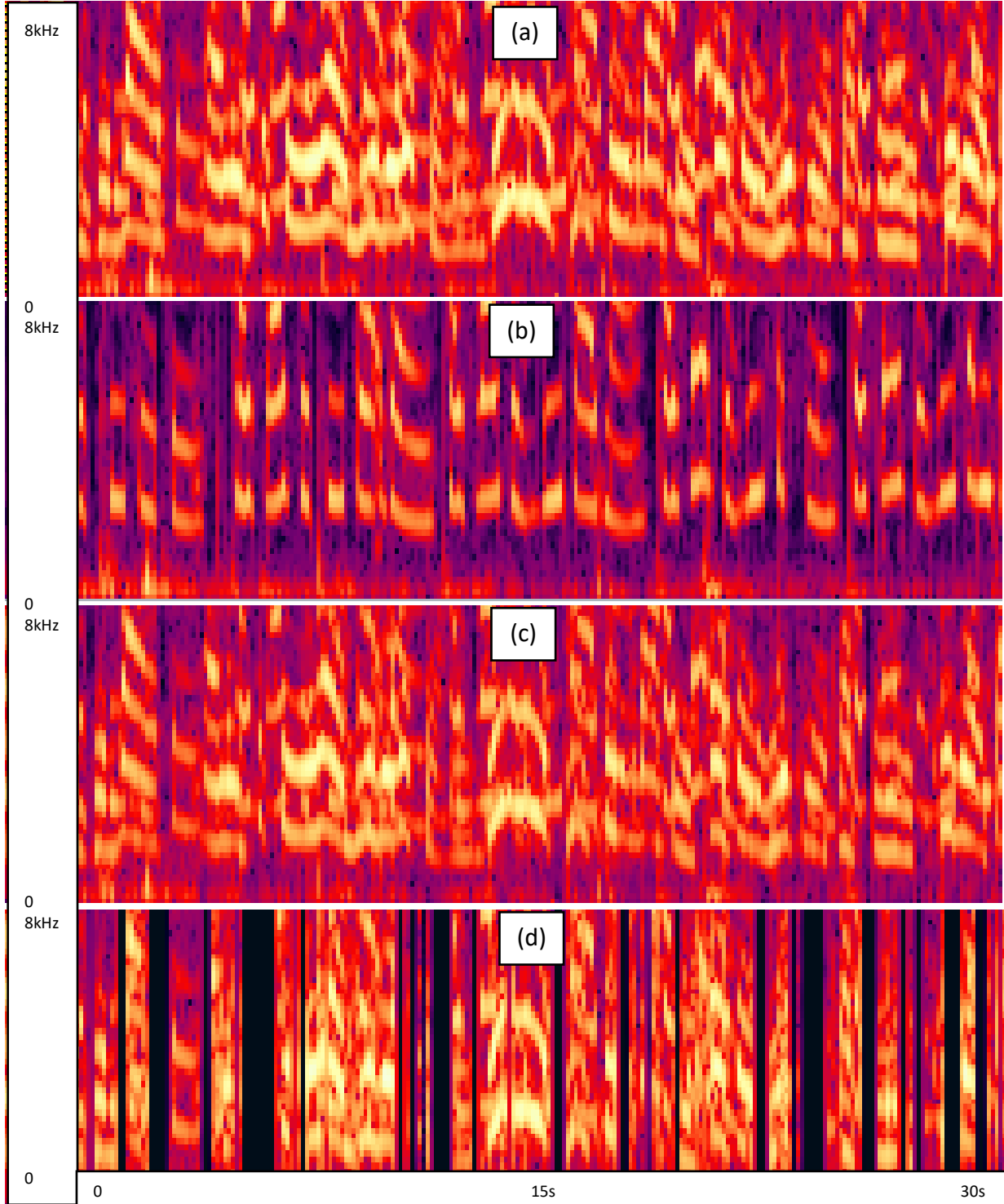


Figure 5.6 Spectrograms of mixture, targeted and recovered speech of the 1st speaker ($f_s=8\text{kHz}$). She is the F of TIMIT. The (a) is the mixture speech. The (b) is the targeted-speech. The (c) is the recovered-speech using the soft mask. The (d) is the recovered-speech using the binary mask. There are horizontal-axes time-domain relationships between all the sketches.

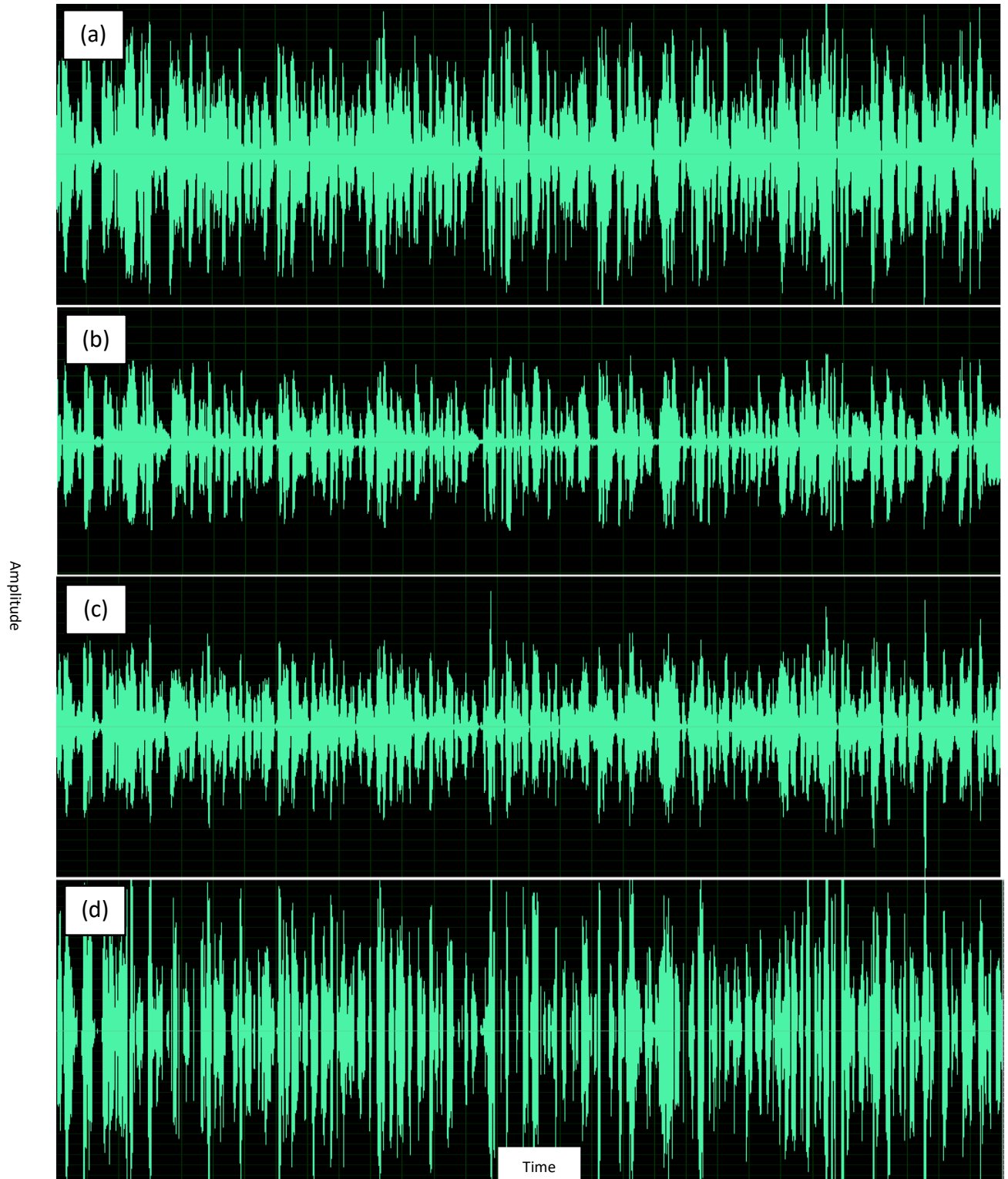


Figure 5.7 Waveforms of mixture, targeted and recovered speech of the 2nd speaker. He is M arbitrary narrator. The (a) is the mixture speech. The (b) is the targeted-speech. The (c) is the recovered-speech using the soft mask. The (d) is the recovered-speech using the binary mask. There are horizontal-axes time-domain relationships between all the sketches.

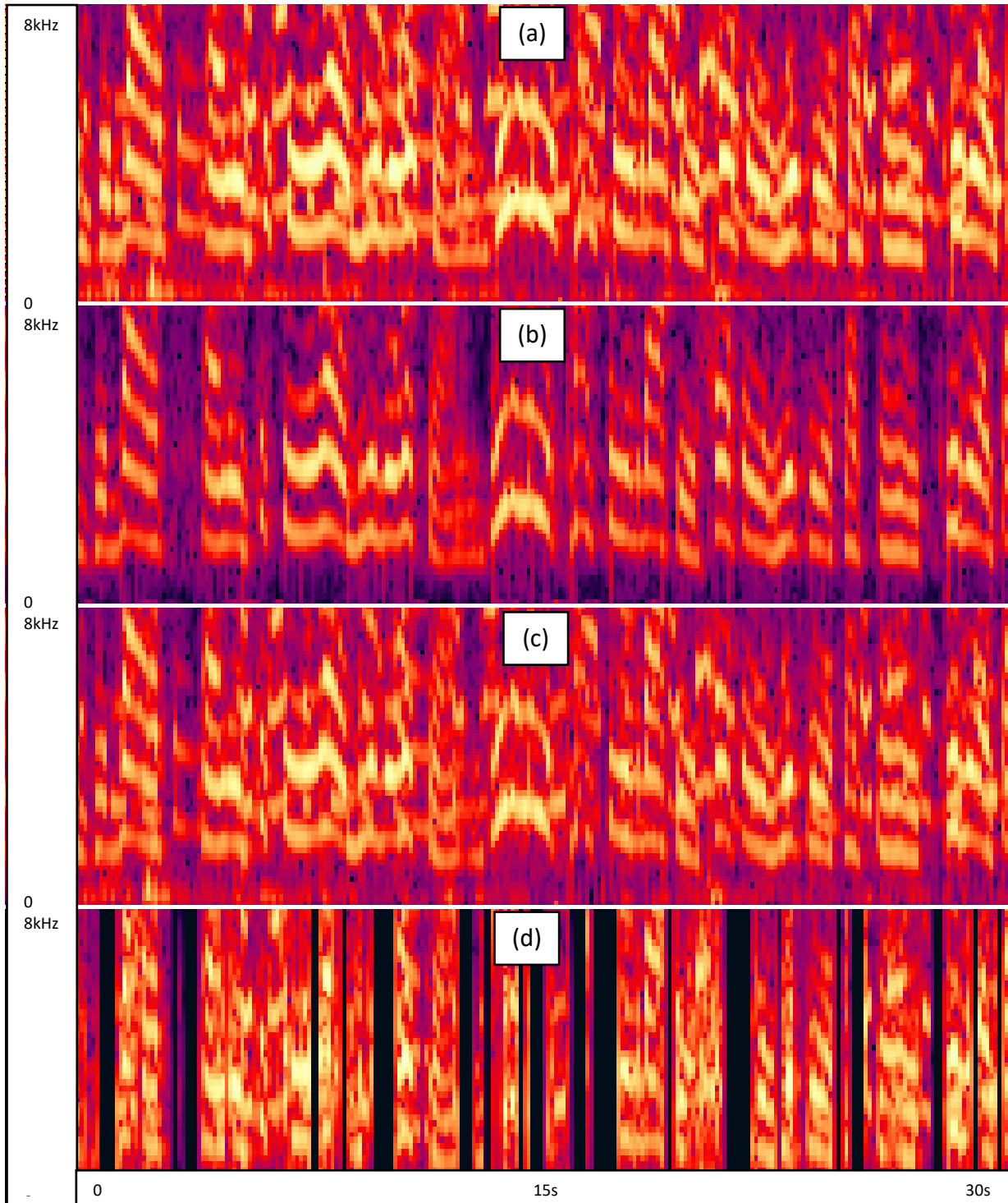


Figure 5.8 Spectrograms of mixture, targeted and recovered speech of the 2nd speaker ($f_s=8\text{kHz}$). He is M arbitrary narrator. The (a) is the mixture speech. The (b) is the targeted-speech. The (c) is the recovered-speech using the soft mask. The (d) is the recovered-speech using the binary mask. There are horizontal-axes time-domain relationships between all the sketches.

- The 4-by-341 files have been checked by the SAR objective test. The results of that test are plotted in Figure 5.9. The upper line graph of the line/Figure 5.9 illustrates the soft masking results of the SAR test (the green graph). The (b)/Figure 5.9 illustrates the binary masking results of the SAR test (the red graph). The (c)/Figure 5.9 illustrates the soft masking results and the binary masking of the SAR test (the green and the red graphs respectively). The (d)/Figure 5.9 illustrates the best optimizes results of the SAR test (the black graph). All these resulting SAR tests are collected in the (d)/Figure 5.9, but the optimizes (black graph) is added by 0.2dB to discriminate it from other tests.

From these 4×341 resulting speech signals, the details of the SAR tests are tabulated in Table 5.1. The table contains: the maximum, the minimum, the average and the variance of the soft, the binary and the optimized masks. The table details the conversations of Female with Male (FM), the conversations of Female with Female (FF), the conversations of Male with Male (MM) and for all the conversations (All).

Table 5.1 SAR objective tests (dB) of Chapter 5 algorithm. The tests for Female with Male (FM), Female with Female (FF), Male with Male (MM) and the all (All) conversations. For each conversation, the Soft and the Binary masks are listed. The optimized results are listed. The table contains: the maximum, the minimum, the average and the variance. The tests are the 1st speaker.

Gender	Mask	Minimum SAR	Maximum SAR	Average SAR	Variance
FM	Soft Mask	7.36	12.07	9.41	1.13
	Binary Mask	4.75	9.53	7.61	0.93
	The optimized	7.96	12.07	9.52	0.92
FF	Soft Mask	4.51	11.45	8.19	1.9
	Binary Mask	5.87	12.69	9.17	2.19
	The optimized	7.45	12.69	9.83	0.99
MM	Soft Mask	5.10	11.21	8.26	1.76
	Binary Mask	5.02	12.30	8.33	2.06
	The optimized	6.85	12.30	9.27	1.08
All	Soft Mask	4.51	11.45	8.13	1.59
	Binary Mask	5.02	12.69	8.94	1.99
	The optimized	6.85	12.69	9.55	1.01

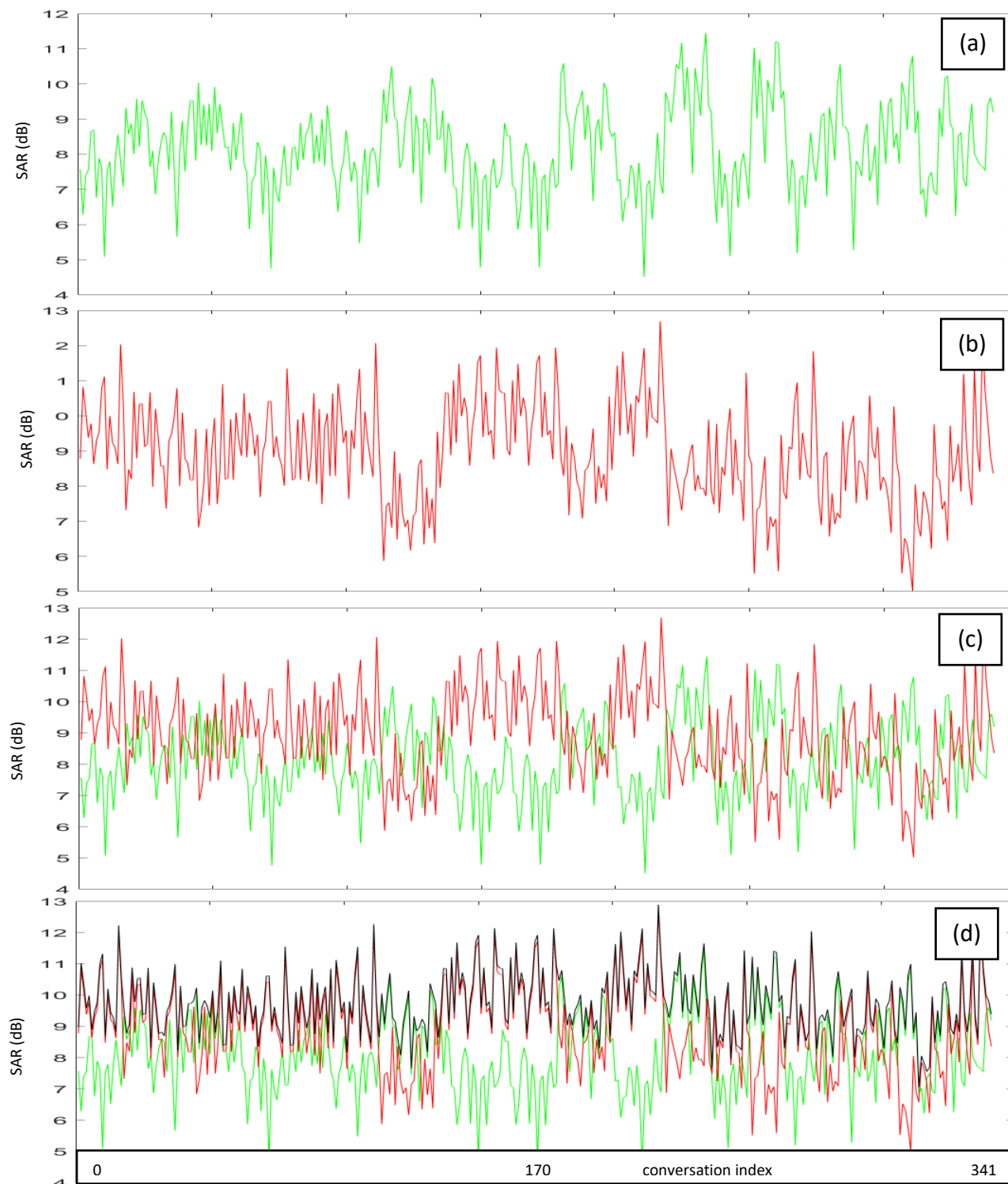


Figure 5.9 SAR tests (dB) of the Chapter 5 algorithm for all the 341 conversations. The (a) (green) using the soft mask. The (b) (red) using the binary mask. The (c) using the soft and the binary masks. The (d) using: the soft, the binary and the best masks, (black) by the optimization. 0.2 dB is added to the best mask for the discriminating between them.

- The 4-by-341 files have been checked by the SDR objective test. The results of that test are plotted in Figure 5.10. The upper line graph of the line/figure 5.10 illustrates the soft masking results of the SDR test (the green graph). The (b)/figure 5.10 illustrates the binary masking results of the SDR test (the green graph). The (c)/figure 5.10 illustrates the soft masking results and the binary masking of the SDR test (the green and the blue graphs respectively). The (d)/figure 5.10 illustrates the best optimized results of the SDR test (the black graph). All these resulting SDR tests are collected in the (d)/figure 5.10, but the optimized (black graph) is added by 0.2 dB to discriminate it from the other tests.

From these 4×341 resulting speech signals, the details of the SDR tests are tabulated in Table 5.2. The table contains: the maximum, the minimum, the average and the variance of the soft, the binary and the optimized masks. The table details the conversations of Female with Male (FM), the conversations of Female with Female (FF), the conversations of Male with Male (MM) and for all the conversations (All). These SDR objective tests are for the first speaker speech signals.

Table 5.2 SDR objective tests (dB) of the Chapter 5 algorithm. The tests for Female with Male (FM), Female with Female (FF), Male with Male (MM) and the all (All) conversations. For each conversation, the Soft and the Binary masks are listed in the table. The optimized test results are listed. The table contains: the maximum, the minimum, the average and the variance values of these collections of tests. These tests are for the 1st speaker tests.

Gender	Mask	Minimum SDR	Maximum SDR	Average SDR	Variance
FM	Soft Mask	-1.67	4.78	1.30	2.67
	Binary Mask	-7.79	-0.44	-3.96	4.24
	The optimized	-1.25	4.78	1.35	2.43
FF	Soft Mask	-6.66	4.30	-0.91	6.38
	Binary Mask	-7.00	4.12	-1.74	6.99
	The optimized	-1.61	4.30	0.78	2.09
MM	Soft Mask	-6.96	5.07	-1.01	7.29
	Binary Mask	-8.23	2.52	-1.80	7.22
	The optimized	-1.62	5.07	0.71	2.32
All	Soft Mask	-6.96	5.07	-0.33	6.62
	Binary Mask	-8.23	4.12	-2.33	7.14
	The optimized	-1.62	5.07	0.91	2.38

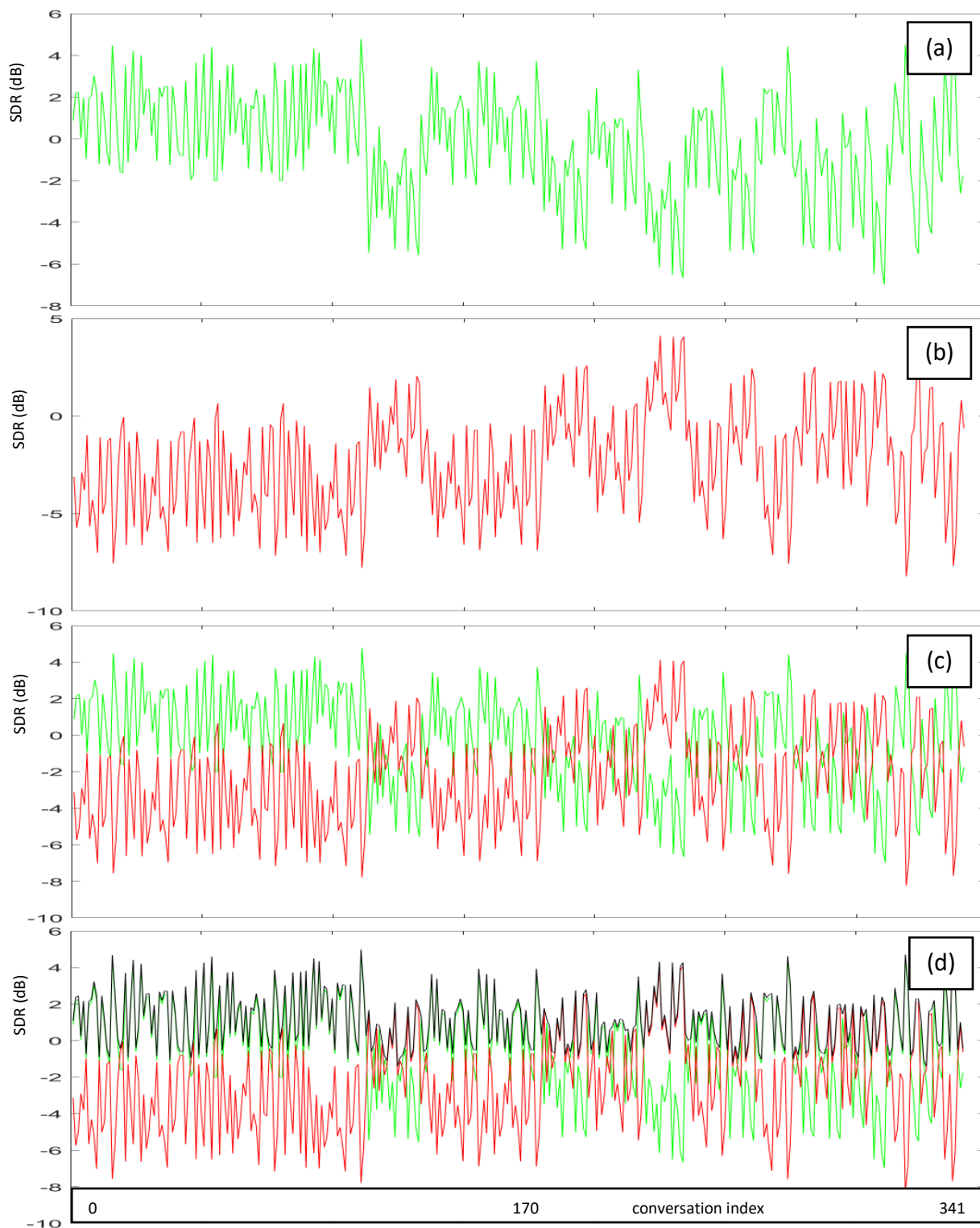


Figure 5.10 SDR tests (dB) of Chapter 5 algorithm for all the 341 conversations. The (a) (green) using the soft mask. The (b) (red) using the binary mask. The (c) using the soft and the binary masks. The (d) using: the soft, the binary and the best masks, (black) by the optimization. 0.2 dB is added to the best mask for the discriminating between them.

- The 4-by-341 files have been checked by the SIR objective test. The results of that test are plotted in Figure 5.11. The upper line graph of the Figure 5.11 illustrates the soft masking results of the SIR test (the green graph). The (b)/Figure 5.11 illustrates the binary masking results of the SIR test (the green graph). The (c)/Figure 5.11 illustrates the soft masking results and the binary masking of the SIR test (the green and the blue graphs respectively). The (d)/Figure 5.11 illustrates the best optimizes results of the SIR test (the black graph). All these resulting SIR tests are collected in the (d)/Figure 5.11, but the optimizes (black graph) is added by 0.2dB to discriminate it from the other tests.

From these 4×341 resulting speech signals, the details of the SIR tests are tabulated in Table 5.3. The table contains: the maximum, the minimum, the average and the variance of the soft, the binary and the optimized masks. The table details the conversations of Female with Male (FM), the conversations of Female with Female (FF), the conversations of Male with Male (MM) and for all the conversations (All). These SIR objective tests are for the first speaker speech signals.

Table 5.3 SIR objective tests (dB) of the Chapter 5 algorithm. The tests for Female with Male (FM), Female with Female (FF), Male with Male (MM) and the all (All) conversations. For each conversation, the Soft and the Binary masks are listed in the table. The optimized test results are listed. The table contains: the maximum, the minimum, the average and the variance values of these collections of tests. These tests are for the 1st speaker tests.

Gender	Mask	Minimum SIR	Maximum SIR	Average SIR	Variance
FM	Soft Mask	-6.00	8.34	3.38	5.49
	Binary Mask	-7.48	0.81	-3.20	5.43
	The optimized	-0.08	8.34	3.44	5.10
FF	Soft Mask	-6.30	7.49	0.55	10.58
	Binary Mask	-6.56	7.57	-0.63	10.72
	The optimized	-1.32	7.57	2.60	4.43
MM	Soft Mask	-6.45	9.12	0.46	12.27
	Binary Mask	-7.70	6.15	-0.53	11.21
	The optimized	-0.59	9.12	2.65	5.13
All	Soft Mask	-6.54	9.12	1.27	11.13
	Binary Mask	-7.70	7.75	-1.23	10.6
	The optimized	-1.32	9.12	2.80	4.96

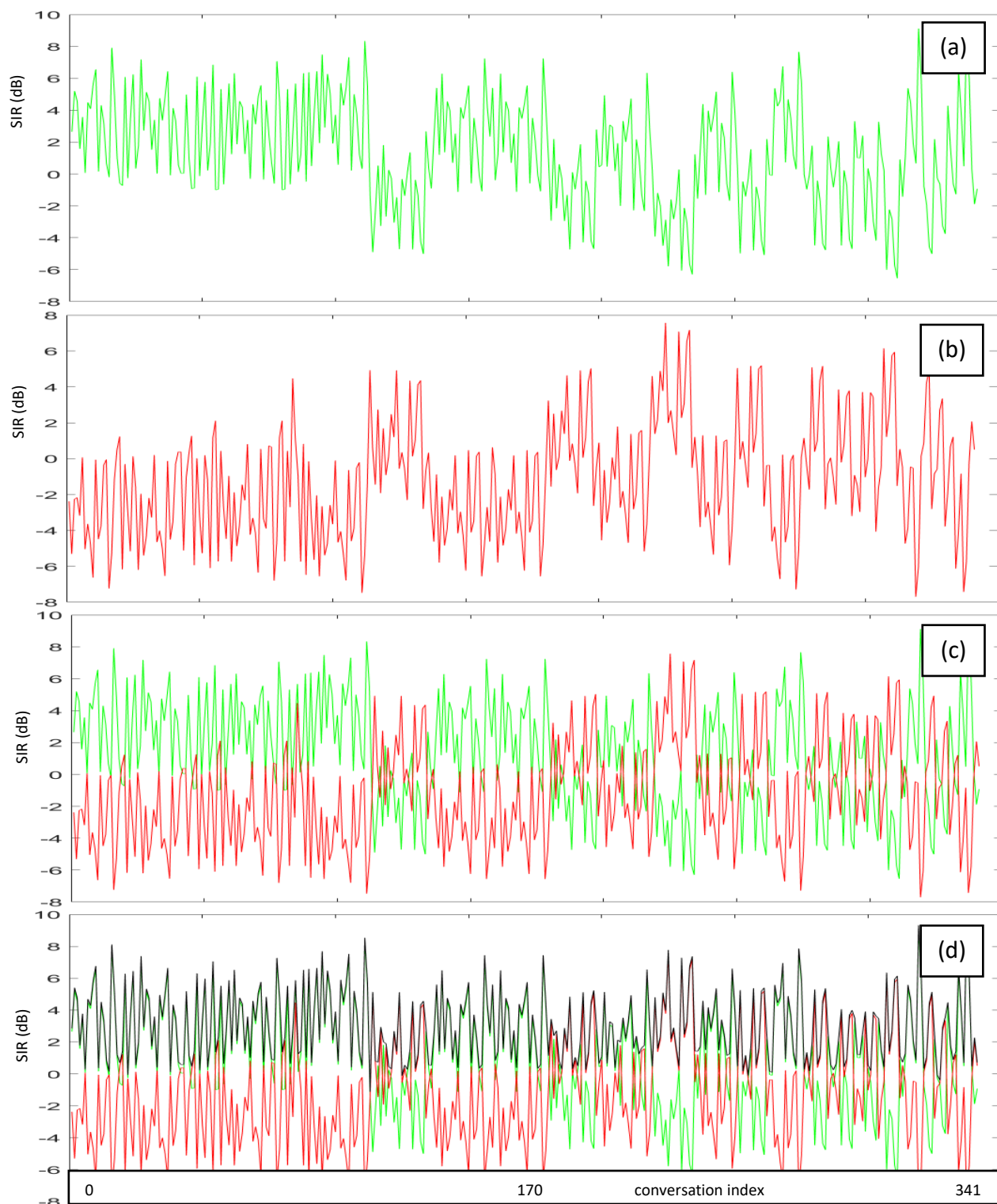


Figure 5.11 SIR tests (dB) of Chapter 5 algorithm for all the 341 conversations. The (a) (green) using the soft mask. The (b) (red) using the binary mask. The (c) using the soft and the binary masks. The (d) using: the soft, the binary and the best masks, (black) by the optimization. 0.2 dB is added to the best mask for the discriminating between them.

For the first speaker, the SAR, the SDR and the SIR tests for all the 341 conversations with their averages are plotted in Figure 5.12. For the second speaker, the average SAR, SDR and SIR tests for all the 341 conversations are tabulated in Table 5.4 and plotted in Figure 5.13. Comparison between the three tests for each speaker, denotes that the tests evaluation is fluctuated from test to other tests. For each one of the two speakers, the three tests values, also fluctuate from speaker to speaker and from test to other tests. Although these evaluations fluctuate, but the average optimized value of each test is approximately equals the average optimized value of another speaker for that test. For the SAR test, the first speaker is 9.55 dB and the second speaker is 9.26 dB. For the SDR test, the first speaker is 0.91 dB and the second speaker is 1.12 dB. For the SIR test, the first speaker is 2.80 dB and the second speaker is 2.97 dB.

Table 5.4 The 2nd speaker objective tests of Chapter 5 algorithm. The tests for Female with Male (FM), Female with Female (FF), Male with Male (MM) and the all (All) conversations. For each conversation, the Soft and the Binary masks are listed in the table. The optimized test results are listed. The table contains: the maximum, the minimum, the average and the variance values.

Gender	Mask	Average SAR	Average SDR	Average SIR
FM	Soft Mask	7.30	1.51	3.15
	Binary Mask	9.17	-3.71	-2.69
	The optimized	9.24	1.56	3.33
FF	Soft Mask	7.98	-0.34	1.18
	Binary Mask	8.84	-1.79	-0.63
	The optimized	9.45	1.08	2.91
MM	Soft Mask	7.96	-0.81	0.66
	Binary Mask	8.03	-1.60	-0.33
	The optimized	8.97	0.91	2.85
All	Soft Mask	7.84	-0.12	1.41
	Binary Mask	8.65	-2.12	-0.95
	The optimized	9.26	1.12	2.97

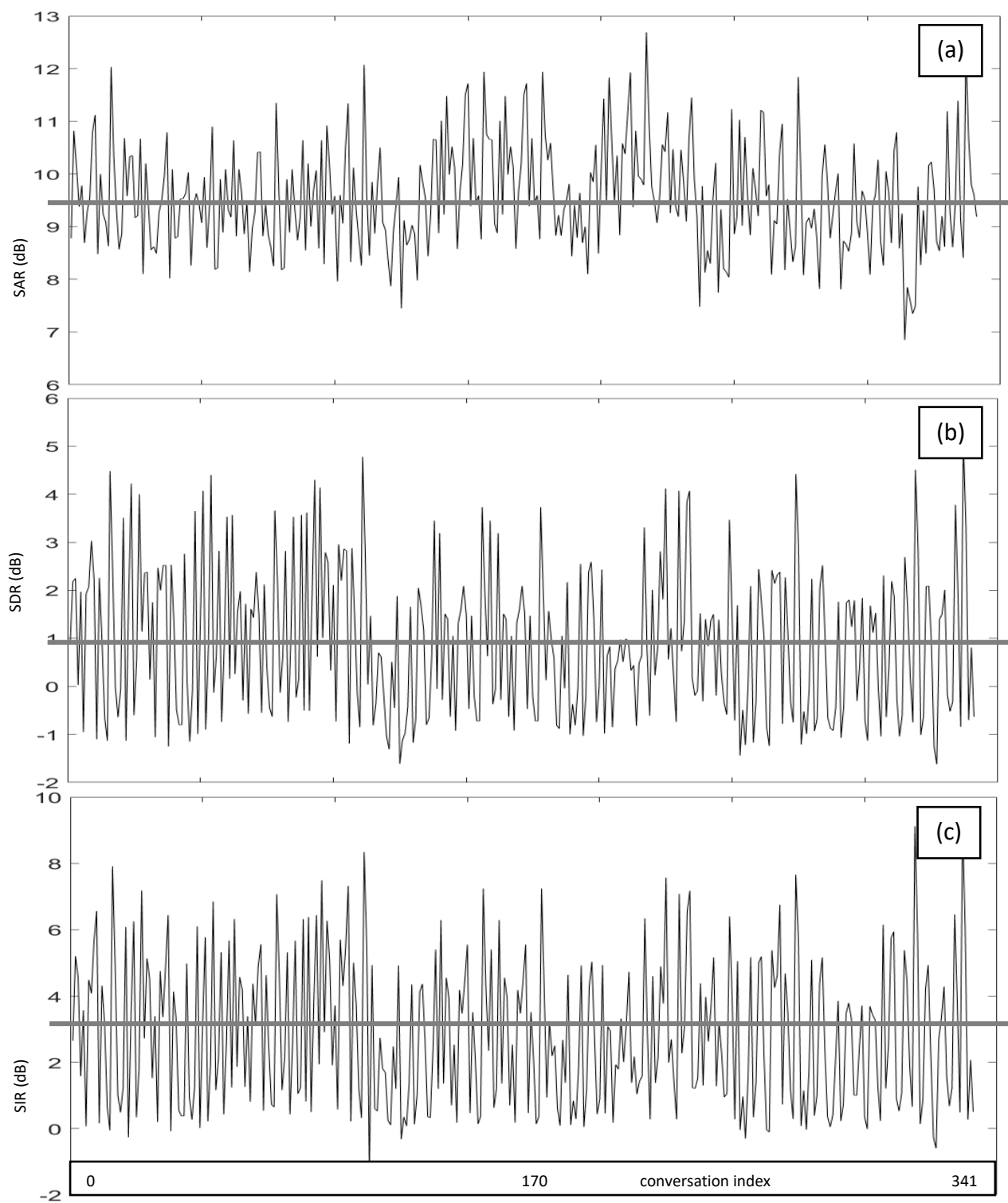


Figure 5.12 The optimized objective tests of Chapter 5 algorithm for 341 conversations. The (a) is the SAR tests. The (b) is the SDR tests. The (c) is the SIR tests. The horizontal line of each plot, is its average value. These tests are for the 1st speaker output speech.

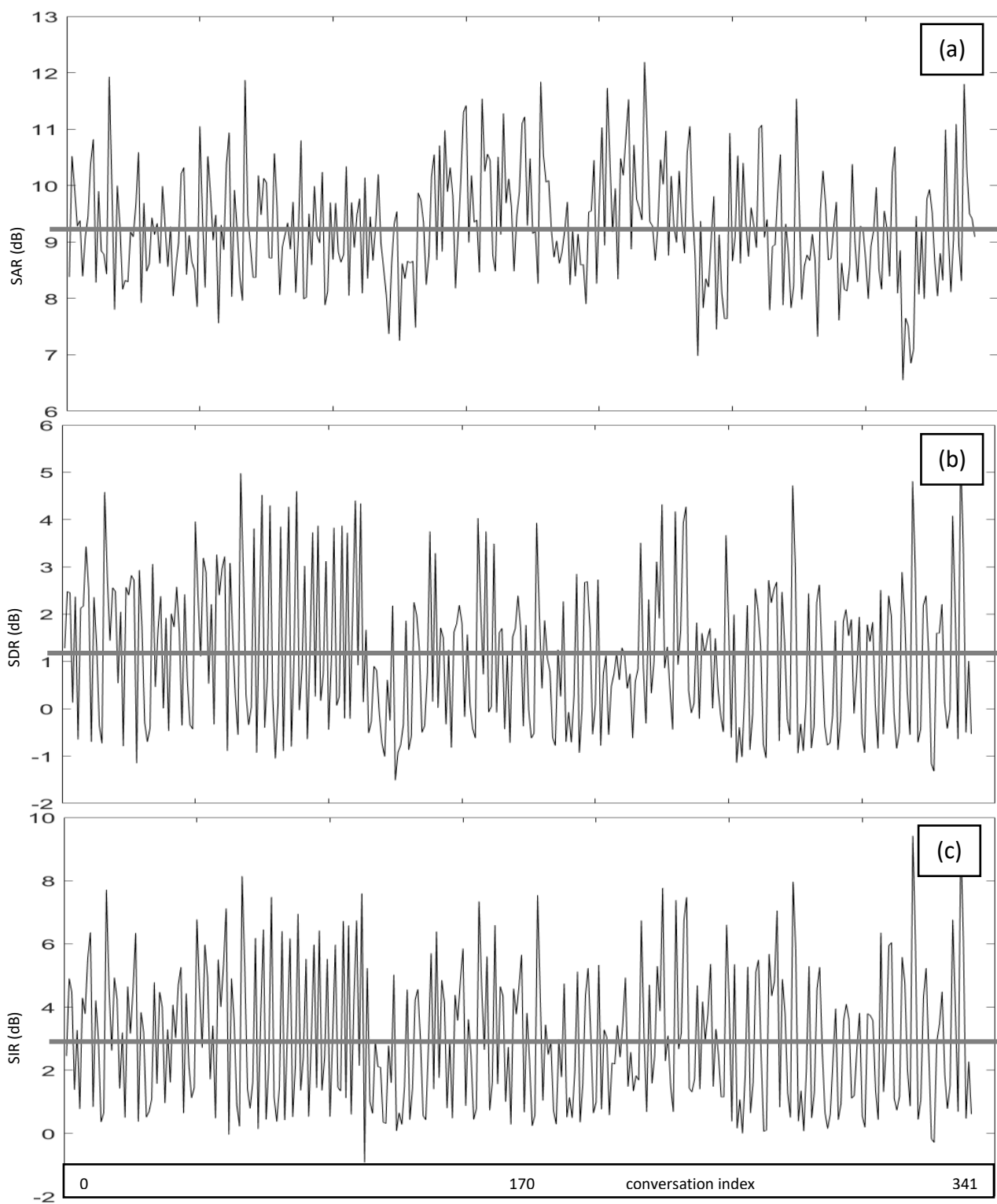


Figure 5.13 The optimized objective tests of Chapter 5 algorithm for 341 conversations. The (a) is the SAR tests. The (b) is the SDR tests. The (c) is the SIR tests. The horizontal line of each plot, is its average value. These tests are the 1st speaker output speech.

For those conversations, average value can evaluate the algorithm. Each test variance assign the robustness of the algorithm against the variations of speakers and speech. Higher variance indicates that the algorithm has sparsity weak-point. Lower variance indicate that the algorithm performance has limited variations for different speakers, genders, subject of speech. According to that merit of variance indication Figure 5.14 and Figure 5.15 illustrate the great change of the optimization for the algorithm performances of the first speaker.

For the first speaker, the gains are: SAR test: The variance gain is $1.59/1.01=1.57$ for the soft and $1.99/1.01=1.97$ for the binary masks. SDR test: The variance gain is $6.62/2.38=2.78$ for the soft and $7.14/2.38=3.0$ for the binary masks. SIR tests: The variance gain is $11.13/4.96=2.24$ for the soft and $10.6/4.96=2.14$ for the binary masks. For the second speaker, the gains are: SAR test: The variance gain is $1.50/1.12=1.34$ for the soft and $1.89/1.12=1.69$ for binary masks. SDR test: The variance gain is $6.52/2.31=2.82$ for the soft and $7.01/2.31=3.03$ for the binary masks. SIR tests: The variance gain is $11.0/4.95=2.22$ for the soft and $10.51/4.95=2.12$ for the binary masks.

From these results, the sparsity of the optimized tests has been reduced by 1.5 to 3 times of the sparsity of the original soft and binary masks. The results of the maximum, minimum and average values of these tests for the first speaker during the 341 spontaneous conversations are bar-plotted in Figure 5.15.

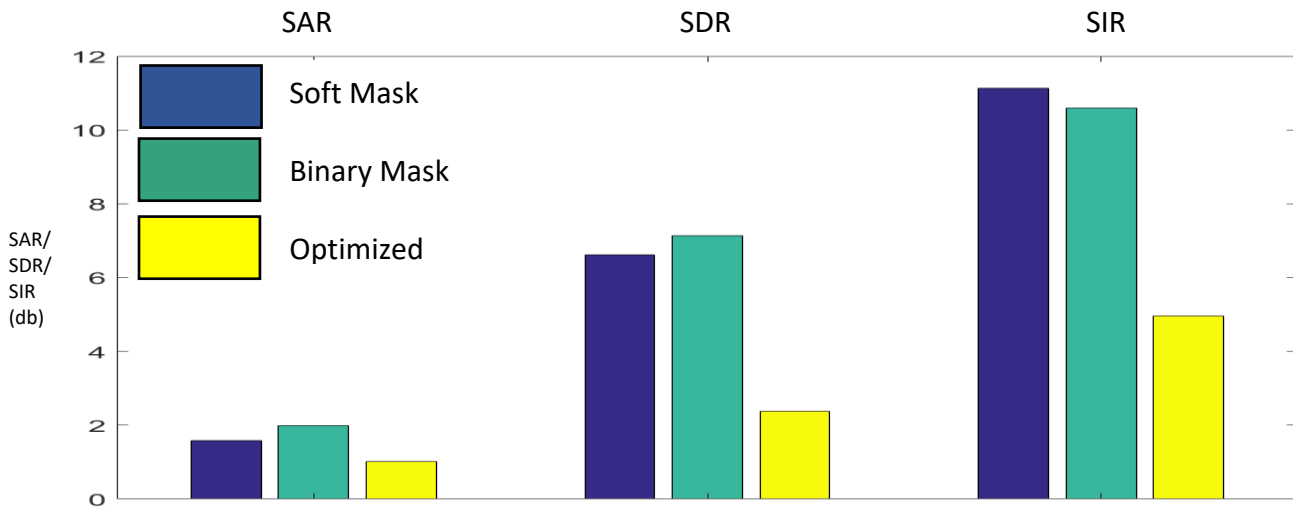


Figure 5.14 Variances of data of the SAR, SDR and SIR 341 tests. The Left-Hand-Side bars are the SAR tests. The middle-collection bars are the SDR tests. The Right-Hand-Side bars are the SIR tests. The blue bars are the soft-masking. The green bars are the binary-masking. The yellow bars are the optimized. These data belong to the first speaker output separated speech signals.

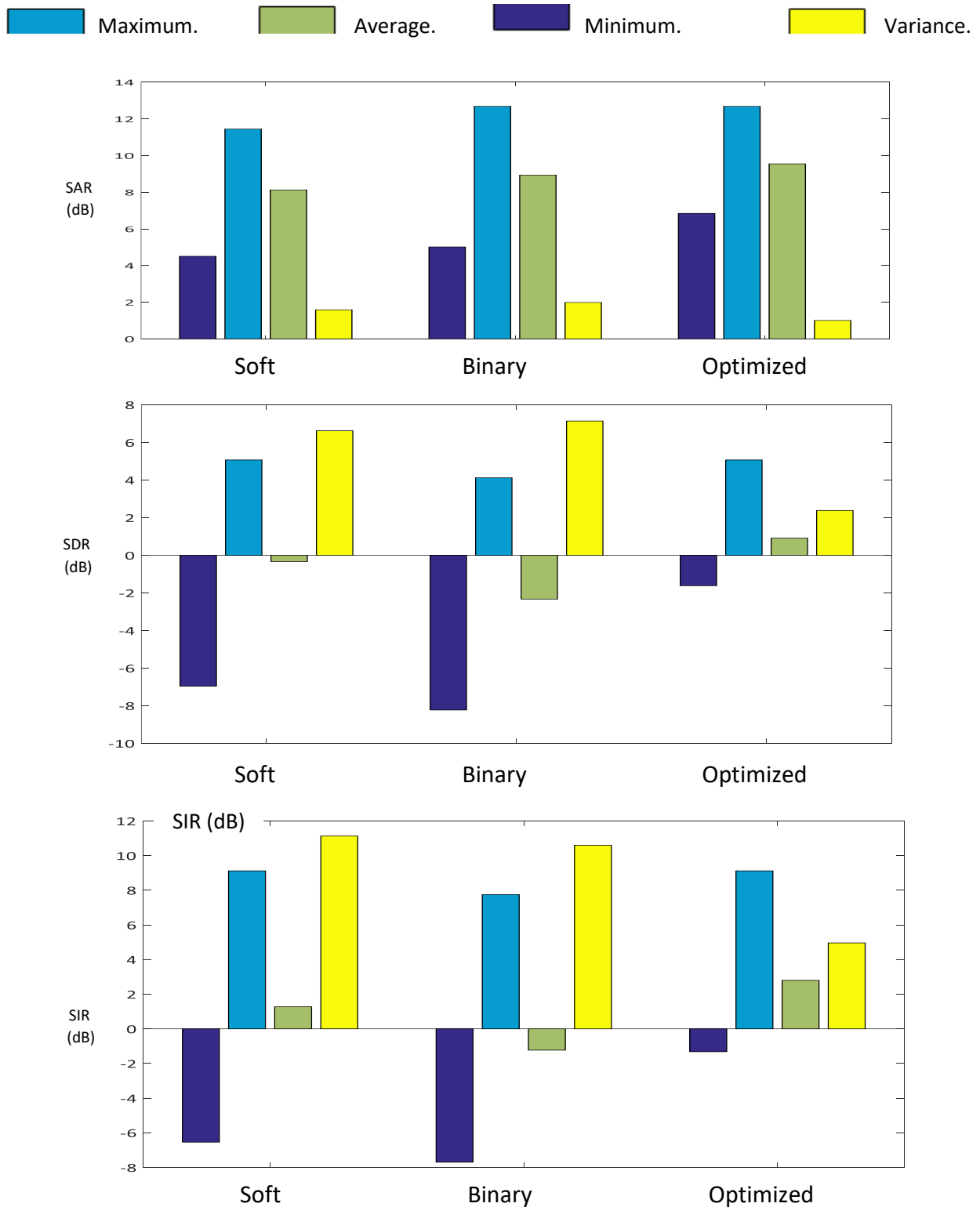


Figure 5.15 Maximum, average, minimum and variance values of the tests. The upper line is the SAR, the middle is the SDR and the lower is the SIR. LHS are the test using the soft masks. Middle are the test using the binary. The RHS are the optimized. These tests are the first speaker output speech.

5.6 Comparison

The resulting tests are compared with recent well-known references. The comparison is listed in Table 5.5 and bar-plotted in Figure 5.16. According to that, the algorithm performance is accepted because it is in the mid-range of these tests. For all the tests, although the minimal/maximal level of this chapter algorithm is less than the minimal/maximal levels of several articles, the minimal/maximal level of this algorithm is larger than the minimal/maximal levels of several articles.

Table 5.5 Comparison with recent well-known articles. The ten references are the highest objective tests. This chapter algorithm is the last row in the table.

	Researchers	Ref.	SAR	SDR	SIR
1	S U Wood & et al.	[168]	4.88 to 7.48	2.84 to 4.16	1.57 to 10.23
2	T Ming & et al.	[111]	15.0 to 15.3	2.9 to 3.0	1.2 to 1.6
3	Y Salaün & et al.	[169]	2.30 to 4.48	-2.49 to -0.91	-5.90 to -1.80
4	C Joder & et al.	[112]	9.8 to 10.4	0.8 to 6.3	0.8 to 10.9
5	A Ozerov & et al.	[170]	4.74 to 10.0	1.52 to 6.62	3.07 to 12.53
6	J Fritsch & et al.	[118]	7.71 to 12.89	5.81 to 10.27	11.88 to 16.62
7	K Adiloğlu & et al.	[171]	5.11 to 9.87	2.65 to 5.57	2.85 to 11.21
8	C Joder & et al.	[172]	8.0 to 11.5	2.0 to 4.5	5.0 to 8.0
9	N Ono & et al.	[173]	5.11 to 9.87	7.73 to 10.25	2.85 to 11.21
10	Z Rafii & et al.	[174]	5.59 to 12.33	0.06 to 10.12	1.30 to 14.66
11	S Arberet & et al. Magoarou & et al.	[175] [176]	4.60 to 12.94	-1.93 to 9.27	-0.74 to 11.98
12	Magron & et al.	[166]	11.0 to 18.7	5.3 to 17	2.8 to 5.3
13	L Wang & et al.	[177]	9.22 to 14.74	6.75 to 12.93	16.69 to 23.27
14	H Kadhim L Woo & S Dlay	[23, 117]	6.85 to 12.69	-1.62 to 5.07	-1.32 to 9.12

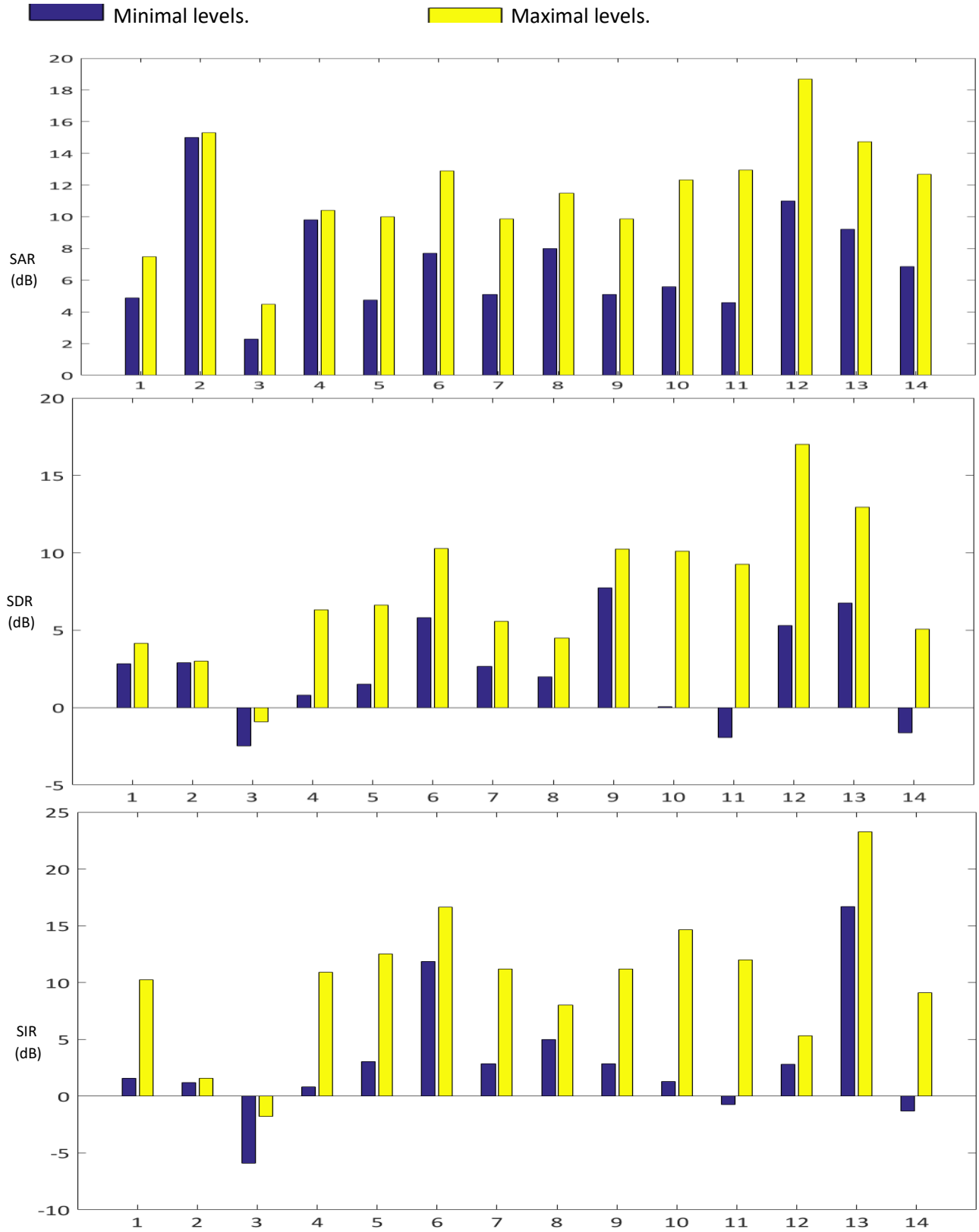


Figure 5.16 comparison with recent tests of well-known articles. The sequence from 1 to 9 is the sequence of Table 5.5.

5.7 Summery

This chapter algorithm and/or Chapter 4 algorithm are the last processing to overcome the main speech-DSP methodologies for the spontaneous conversations. Chapter 3 has segregated the conversation signal into two speech format signals: the dialogue and the mixture. The research did not focus on the dialogue speech processing which is the speaker diarization. The research focused on the mixture speech signal. To overcome the mixture speech problem, the research has approaches to speech separation using its two branches: the blind and the informed. In Chapter 4, the blind speech separation is achieved.

In this chapter, the process is informed speech separation, where the main observation input signal is the mixture speech which is the output of the chapter 3 overlapped-speech detection. The dialogue should be processed by the speaker diarization to produce individual speech of each speaker. Those speech signal is the database-like for the separation process.

In this chapter, the database-like signals are supposed as virtual individual speech and their summation is virtual mixture speech. To discriminate between these signals and the input mixture, the input is called real mixture speech signal. The virtual signals are trained using the analogy between the NMF matrices of the real and the virtual speech. That assistance of the database-like easy the processing job. Because the quality of the resulting separated speech fluctuates depending on the masking type, both soft and binary masks are used. The choice of the best mask is based-on the objective tests: SAR, SDR and SIR.

The informed speech separation of this chapter is moderate when it compared with the recent standard well-known articles. The range of the tests for this chapter algorithm is in the middle range among those article achievements.

Chapter 6. Notes, Conclusions and Future Works

6.1 Notes and Conclusions

Through the Chapter 3, 4 and 5 processes, the following notes and conclusions are remarked:

- There is no significant effect of the sampling rate and the sample resolution, neither on the overlapped-speech detection nor on the speech separation. Both the 8000 and the 16000 samples/ second sampling rates are used successfully for those. The effect of 32000 samples/ second sampling rate does not provide any extra efficiency for the performance, so it is neglected. The research uses 16 bit/ sample for the resolution. The 24 bit/ sample and the 32 bit/ sample do not increase the audio definition significantly. The speech is deterministic by the using of 16 bit/ sample resolution.
- For the two speakers gender of the spontaneous conversation, the resulting subjective and objective tests, denote that the gender has limited effect on those tests. Each simulated spontaneous conversation is talking about specific subject(s), because each conversation is audio book. Ordinary books are focusing on known subjects. In the research, the conversation subject does not have clear effects.
- The performance of the overlapped-speech detection algorithm is excellence. The errors are only near the switching instances. For the subjective testers, the errors almost are not sensible. At the worst cases, the errors are audible for less than one second. The subjective tests ensure that the performance is excellent compared with the standard speaker diarization corpuses. The silent period inside of the conversation should be considered. The performance of the algorithm equals the performance of several current corpuses, and better than other corpuses.
- Although the performance of the blind speech separation is moderate, it is very good compared with NMF speech separation. According to the current literatures, NMF technique has good ability for the audio separation, but it has bad ability for the speech separation.
- The performance of the informed speech separation depends on two factors. The first is the know-how of the matrix-configuration for the observation signals and the generated database. The second factor is the switching step for the optimization to choose the highest objective tests. The performance fluctuates from very low to very high. The optimization

mitigates that wide-range fluctuating. The statistical properties of the resulting speech prove that the optimization reduce the fluctuating range.

- Type of the speech separation mask is important for the efficient process. The weak-point of the binary hard mask is the discontinuity of the output separated speech. There are two methods for the soft masking, the amplitude and the power calculations.
- Since the speech is quasi-stationary signal, all its parameters (e.g. audio features) are dynamically changing from speaker to speaker, from speech to speech, from sentence to sentence. For specific speaker, that merit is shown for speaking specific sentence on different conditions.

6.2 Future Works

The following are the suggested future works for the overall system, the chapter 3 overlapped-speech detection algorithm, the chapter 4 blind speech separation algorithm and the chapter 5 informed speech separation algorithm.

For the overall system:

- Because the observation speech signal of the research is 2-speaker spontaneous conversations, the overall system could be repeated with more than two speakers.
- Number of speakers is known in the chapter 3 algorithm. The algorithm could be repeated with unknown number of speakers. This case is more challenge against the researchers, but this case should be considered.
- The system is either unsupervised or semi-supervised system. Partial or complete information/ database for part or all the speakers, change the system to supervised ML. That change gives the researcher(s) extra tools/ facilities to achieve the objectives and the aims of the research.
- In this thesis, the processing sequence is: the overlapped-speech detection, the speaker diarization then the speech separation. The sequence may be changed to diarization then detection, or may be done by the parallel processing of two processes at specific time then the third process.

For the chapter 3 overlapped-speech detection algorithm:

- The MFCC, the LFCC, the LPC or the PNCC features may be used instead of the RASTA-PLP features.

- Instead of the 13 features per frame, 18, 24, 32 or 40 increases the definition of those parameters.
- For the group concept, number of features per fundamental group covers other duration instead of the 0.1 second.
- Beside the principles of the pattern recognition, GMM and HMM models can optimize the features efficiently.
- Other thresholds of the scenarios are important for better speaker clustering of the speaker segments. Feedback of the subjective tests to the detection could improve its performance.

For the chapter 4 blind speech separation algorithm:

- Other techniques (e.g. ICA, PCA, CASA) can separate the mixture output of the input conversation.
- For the NMF, changing the other parameters of the factorization gives chances for other results.
- In Chapter 4, number of sub-signals are 24 per Filter-Bank sub-band. The long unfeasible run-time to simulate the algorithm, enforce to reduce that number. Short time run-time admits to increase that factorization parameter. The increasing expands the resolution of the filter-bank/ NMF analysis and the speech separation.
- The choosing of the efficient speech separation mask, statistical study helps help the researchers to use specific mask and omit the other.

For the chapter 5 informed speech separation algorithm:

- The kernel of the algorithm is the configuration of the virtual-speech with the real-speech matrices. Trial-and-Error of other configurations expand the chances for higher SAR, SDR and SIR subjective tests.
- Dimensions of these matrices also, expand these chances.
- The positive-element condition of the NMF matrices are omitted in the Chapter 5. The condition regardless because there is no another choice to insert the database-information into the separation process. The existence of that condition might increase the speech separability.
- The optimization step for choosing the better separated speech, needs more statistical study. In the research, only the variances of the optimization choices are calculated. More statistical details might help the researchers to choose the better mas.

Appendix A. Historical Overviews

A.1 Historical Overview of Filter-Bank

Digital Filter-Bank has the same principle of the traditional analogue filter-bank. There are a lot of themes in touch with filter-bank: implementation, tasks and speech compression. At the beginning era of DSP, the researchers were taking care on implement digital filters in the time domain by a minimal number of iterations. Implementation time of the digital filter is the main reason for reducing the iteration number. Also, filter-bank researchers were taking cares on the same challenges and their solutions. The famous researchers who had achieved: the principles, the theorems, the algorithms, the implementation and the applications of the digital filter-bank are: *S Stevens, J Volkman and D Gabor* during the 1940s. *D W Robinson, B Smith, M R Schroeder, E E David and R S Dadson* during the 1950s. *B Gold, C M Rader, J L Flanagan, R M Golden, A M Noll, A V Oppenheim, R W Schafer, L R Rabiner, A E Rosenberg, K Ishizaka, J L Flanagan, M D Paez and T H Glisson* during the 1960s. *M R Sambur, C K Un, D T Magill, V Viswanathan, J Makhoul, P Noll, M R Portnoff, J D Markel, A H Gray, V Viswanathan, R Schwarz and A W F Huggins* during the 1970s.

A.2 Historical Overview of k-means

Clustering algorithms are the most important methodologies to perform machine learning and pattern recognition jobs. The main task of the clustering is the creation of proper borders (thresholds) to divide specific data into (known or unknown) number of clusters. The clustering algorithms are based-on: the connectivity between the data (the hierarchical clustering), the centroid of clustered-data (e.g. k-means), the distribution of the data and the density of the data. The simplest and the most popular clustering method is the k-means algorithm. The algorithm of k-means is iterative looping. k-means clustering can divide any known-number of data into k known-number of clusters. At first, k-means algorithm finds the mean (centroid) of each cluster. The other data are belonged to the nearest centroid; hence, each datum is located in the proper cluster. That algorithm is done for iterative loops till the Euclidian distances reach the converging condition.

Historically, k-means term was created by the idea of *Hugo Steinhaus* in 1957. During that year, *Stuart Lloyd* proposed the first (the standard) k-means algorithm. Lloyds did not publish his work at that time, but he did that later in 1982. In 1965 *E W Forgy* wrote details about the same algorithm of Lloyd. *James MacQueen* is the first mathematician used the k-means term in 1967. In 1975 and in 1979, *Hartigan* and *Wong* simulated the Lloyd-Forgy algorithm by an efficient PC programing. During the past 60 years, a lot of papers had been published to enhance the standard Lloyd algorithm by proposing the modification algorithms. The famous authors of these papers are: *Mardia, K V, Gordon A D, Hubert L J and Everitt B S* during the 1970s, *Wong M A, Spath H,*

Milligan G W and *Pollard, D* during the 1980s, *Garcia-Escudero L A* during the 1990s, and *Amorim R C* during the 2000s. Through the half century of k-means history, the standard algorithm was encountering two major problems. The first problem is the required time to complete its numerical-analysis calculations. The sequence of the calculations is the finding of the centroids of the clusters then the sharing of the data into their proper clusters. Second problem occurs when any centroid sticks inside one cluster during the iteration loops. This problem causes an ill-condition for the solution, i.e. infinite iterative looping. To avoid these problems, two significant efficient improvements are proposed. The performance of those improvements overcomes these problems. Improvement of *Yuan Zhang, Zhongyang Xiong, Jiali Mao* and *Ling Ou* produce the term k-means || (parallel) in 2006. In 2012, *Alon Vinnikov* and *Shai Shalev-Shwartz* suggested the k-means ++ (plus plus) improvement. k-means || and k-means ++ reduce the required time for the calculations. The improving algorithms do not suffer from any sticking of data inside any cluster.

A.3 Historical Overview of Overlapped-Speech Detection

Speaker diarization splits speech signals of several speakers from their dialogue conversation. The input signal is a conversation where only one speaker is speaking while the other speakers are silent. When that speaker quiets, one silent speaker continues the speaking, and so on. This format of dialogue talking is identical. The spontaneous conversation contains durations of an overlapped-speech of multi-speakers. Any duration of overlapped-speech can be neglected if it does not have regarded information. sometimes, these durations have regarded information. In this case, the overlapped-speech signal is segregated to its original speakers' signals.

O Ghitza discussed the problem of the overlapped-speech by analysing and synthesizing the overlapped-speech. In 1986, he matched the synthesized speech with its representation. In 1987, he repeated that by the auditory nerve representation.

In beginning of the 1990s, the speaker diarization field begun by *M Sugiyama, J Murakami, H Watanabe, C J Leggetter, P C Woodland, U Jain, M A Siegler, S-J Doh, E Gouvea, J Huerta, P J Moreno, B Raj, R M Stern, F Kubala, H Jin, S Matsoukas, L Nguyen, R Schwartz, J Makhoul, K Shinoda, T Watanabe, M J F Gales, D Pye, S J Young, V Parikh, S Chen, P S Gopalakrishnan, R A Gopinath, H Printz, D Kanevsky, P Olsen, L Polymenakos, D Kanevsky, H S M Beigi, S H Maes, A Solomonoff, A Mielke, M Schmidt* and *H Gish*.

In 1992, *Furui* investigated the recognition of the speech under the overlapping conditions. In 1996, *Kobyashi, Kajita, Takeda* and *Itakura* extract the features of the overlapped-speech signal. The first attempts for separating an overlapped-speech were in 1997 and 1998 by *Taniguchi, Kajita, Takeda* and *Itakura*.

A.4 Historical Overview of Speech Separation

Speech separation is part of source separation DSP. During the II Word War, source separation had been analysed by *Colin Cherry*, under the title Cocktail-Party Problem. Through the past seven

decades, speech separation is a challenge which was facing the speech-DSP researchers. Mainly, speech separation isolates the components of the mixture speech. The speech separation tries to recover the original speech (is called the targeted-speech) of each speaker alone. In 1961, *H Barlow* suggested a neural network algorithm to separate signals by the mutual relation between the input and output data. The first productive attempts were in the middle of 1980s. In 1988, the works on blind source separation are by *C Jutten*, *J Herault* and *A Guerin*. In 1989, several papers had been presented and published in the proceeding of the workshop on Higher-Order Spectral Analysis at Vail, NJ. The papers are presented and edited by *H H Chiang*, *C L Nikias*, *P Comon*, *J L Lacoume*, *P Ruiz* and *J F Cardoso*. In 1990, *M Gaeta* and *J Lacoume* proposed an approach that using the maximum likelihood estimation.

On the first half of the 1990s, the most significant related-researches continued by: *J Karhunen*, *J Joutsensalo*, *A Chickocki*, *L Moszczynski*, *J Attik*, *J Nadal*, *N Parga*, *G Burel*, *D Yellin*, *E Weinstein*, *H Ngyuen*, *R Linsker*, *S Becker*, *G E Hinton*, *A J Bell* and *T Sejnowski*.

In 1994, *P Comon* introduced the Independent Component Analysis ICA. I that year, *J Karhunen*, *E Joutsensalo* and *E Oja* introduced the nonlinear Principal Components Analysis PCA.

On the second half of the 1990s, the most significant related-researches continued by: *Z Roth*, *Y Baram*, *B Laheld*, *M Girolami*, *C Fyfe*, *T W Lee*, *M Girolami*, *S Amari*, *B A Pearlmutter*, *L C Parra*, *S Makeig*, *T-P Jung*, *M J McKeown*, *M Barlett*, *M S Gray*, *J Movellan*, *K Torkkola*, *R Orglmeister*, *M Hermann*, *H Yang*, *J K Lin*, *D G Grier*, *J D Cowan*, *P Pajunen*, *B Koehler*, *A Taleb*, *Le Blanc*, *De Leon*, *K C Yen*, *Y Zhao* and *A Cichocki*.

Computational Auditory Scene Analysis CASA is an algorithm used for the speech separation. The algorithm simulates the hearing knowhow of human ear. During the past century, Audio and speech laboratories attempted to measure the characteristics of the human ear. The goal of those measurements is the corresponding simulation of the human ear. In 1996, *D Ellis* derives the details of CASA using those measurements. CASA simulates the capability of the human ear and brain for chasing specific speaker in a multi-speaker conversation.

A.5 Historical Overview of NMF and NMF-based Speech Separation

Historically, in beginning of the 1970s, Non-negative Matrix Factorization NMF was called Self-Modeling Curve Resolution SMCR. In the middle of that decade, the terms Factorization and Non-negative-matrix appeared on literatures of several mathematical researchers, e.g. *A Berman*, *R J Plemmons* and *L B Thomas*. First attempts for factorization of the non-Negative matrix started during the 1980s by *S L Campbell*, *G D Poole*, *R D Paola*, *J P Bazin*, *F Aubry*, *A Aurengo*, *F Cavaillioles*, *J Y Herry*, *E Kahn*, *J S Kahn*, *J Shen* and *G W Israël*.

In the 1990s, early work on the NMF was performed by mathematical researchers. They are the Finnish group: *P Paatero*, *U Tapper*, *A Berman*, *R J Plemmons*, *P Anttila*, and *O Järvinen*. In that decade, *Daniel D Lee* and *H Sebastian Seung* published their papers of NMF, then the term became well-known.

During the first decade of the 2000-Millennium, the NMF had been widely exploited for different numerical analysis and statistical jobs. Large number of literatures were published by: *H Sitek, O Patrik, W Xu, Y Gong, M Rosen-Zvi, G E Hinton, F Årup, B Daniela, L Kai, M W Berry, M Browne, C Ding, X He, H D Simon, R Zass, A Shashua, E Gaussier, C Goutte, S Sra, X Liu, N N Zheng, Q B You, R Kompass, Y Mao, L Saul, J M Smith, A N Langville, V P Pauca, R J Plemmons, C-J Lin, M N Schmidt, J Larsen, F T Hsiao, F Å Nielsen, H Kim, Y Y Tang, N-D Ho, V Blondel, W Peng, C Boutsidis, S-i Amari, N Bertin, J-J Durrieu, T Zhang, B Fang, W Liu, G He, J Wen, P Van-Dooren, K Devarajan, A Cichocki, R Zdunek, V K Potluru, S M Plis, M Morup, V D Calhoun, T Lane, S A Vavasis, R Tandon, T Li, M Jordan, C Liu, H-c Yang, J Fan, L-W He, Y-M Wang, W Wang, J Eggert and E Körner.*

In the 2010s, the researchers continued with their modifications on the NMF e.g. *Y Chen, X Wang, C Shi, E K Lua, X Fu, B Deng, X Li, C J Hsieh, I S Dhillon, R Gemulla, E Nijkamp, P J Haas, Y Sismanis, K Yilmaz, A T Cemgil, U Simsekli, J Kim, H Park, V Kalofolias, E Gallopoulos, N Guan, D Tao, Z Luo, B Yuan, L Taslaman, B Nilsson, S Arora, R Ge, Y Halpern, D Mimno, A Moitra, D Sontag, Y Wu, M Zhu, J Liu, C Wang, J Gao, J Han, E Schwalbe, D Wang, V Ravichander, E Nick, T F Zheng, Y Jiangtao, L Gao, Z (Mark) Zhang, M-Y Kan, P Xie, X Chen, Y Bao, H Fang and J Zhang.*

After the introduction of *Lee and Seung*, the decomposition ability of the NMF is for-and-only-for the positive-element matrices. Applying of that ability has been started several years later. The first attempt, of that application was by *Hoyer* on the Non-negative Spares Coding. He gave a simple efficient multiplicative algorithm to find optimal values of the hidden components. He illustrated how the Basis Vector can be learned from the observed data. His effective proposal method is simulated and demonstrated. On the same field (the Spares Coding), *Eggert and Edgar Korner* showed how to merge concepts of the Non-negative Factorization with sparsity conditions. It is a multiplicative algorithm which is comparable in efficiency to the standard NMF. That can be used to gain sensible solutions in the over-complete cases. The case for learning and modelling the arrays of receptive fields arranged in a Visual Processing Map, where an over-complete representation is unavoidable.

Appendix B. Software

MATLAB is the main software environment to simulate major implementations of the research thesis. It is a powerful reliable application to support most operations of: DSP algorithms, statistical parameters and numerical analysis [178]. In addition to the new proven algorithms, online reliable MATLAB toolboxes, sometimes are available. The toolboxes help scientific researchers for cross-checking their codes. Although the MATLAB is based on the quickest Object Oriented Programming (C++ /OOP), main problem of the MATLAB simulation is the required runtime to complete essential calculations of the simulated experiments. For example, to build a database of audio features for a specific speaker, one hour of his/her speech should be available. To perform such job, several hours are spent for that job of calculations. To reduce the long runtime, efficient edited MATLAB codes can be used. Unfortunately, efficient codes cannot overcome the time problem entirely. Despite that problem, MATLAB is faster than any other high level language environment (e.g. FORTRAN and BASICS). To reduce such long calculation time, structural languages (e.g. C and PASCAL) could be used. The structural languages are faster than MATLAB, but they are more complicated than the MATLAB. Almost, the programmer prefers the simplicity in editing, compiling, running and debugging the written code. For these reasons, most of DSP researchers avoid the structural languages and use the MATLAB environment.

In addition to the simplicity, MATLAB has a lot of tools and facilities. MATLAB can visualize sequence of numerical data by plotting its equivalent curves and bars. Those graphics could be discriminated by using different colors and signs. Figure legend and labels are edited then added on the plots to rich the figures perception [178].

In speech-DSP, MATLAB use to: read the observation signal speech files, read the data-base binary files, write the outputs speech files, manipulate the read and the written arrays, link with other software applications (e.g. LABVIEW), interface with hardware modules (e.g. Arduino). MATLAB has ability to input and output live speech signal during the runtime. Such capability enables the researcher to avoid the read and the write instructions and follow the processed speech signal by hearing the tested tips, directly on the MATLAB host machine.

To simulate the experiments of a research, the researcher could not edit, debug and run the simulation, entirely. The simulation from a scratch is not feasible, because it needs a lot of time. The much time is due to the long runtime, a lot of errors and trial-and-error routines. To reduce that time with less of errors, the existing reliable toolbox of MATLAB codes are invoked. The following are the main reliable online toolboxes (with their URL), which are invoked in the simulation of the experiments of the Chapter 3 algorithm, the Chapter 4 algorithm and the Chapter 5 algorithm:

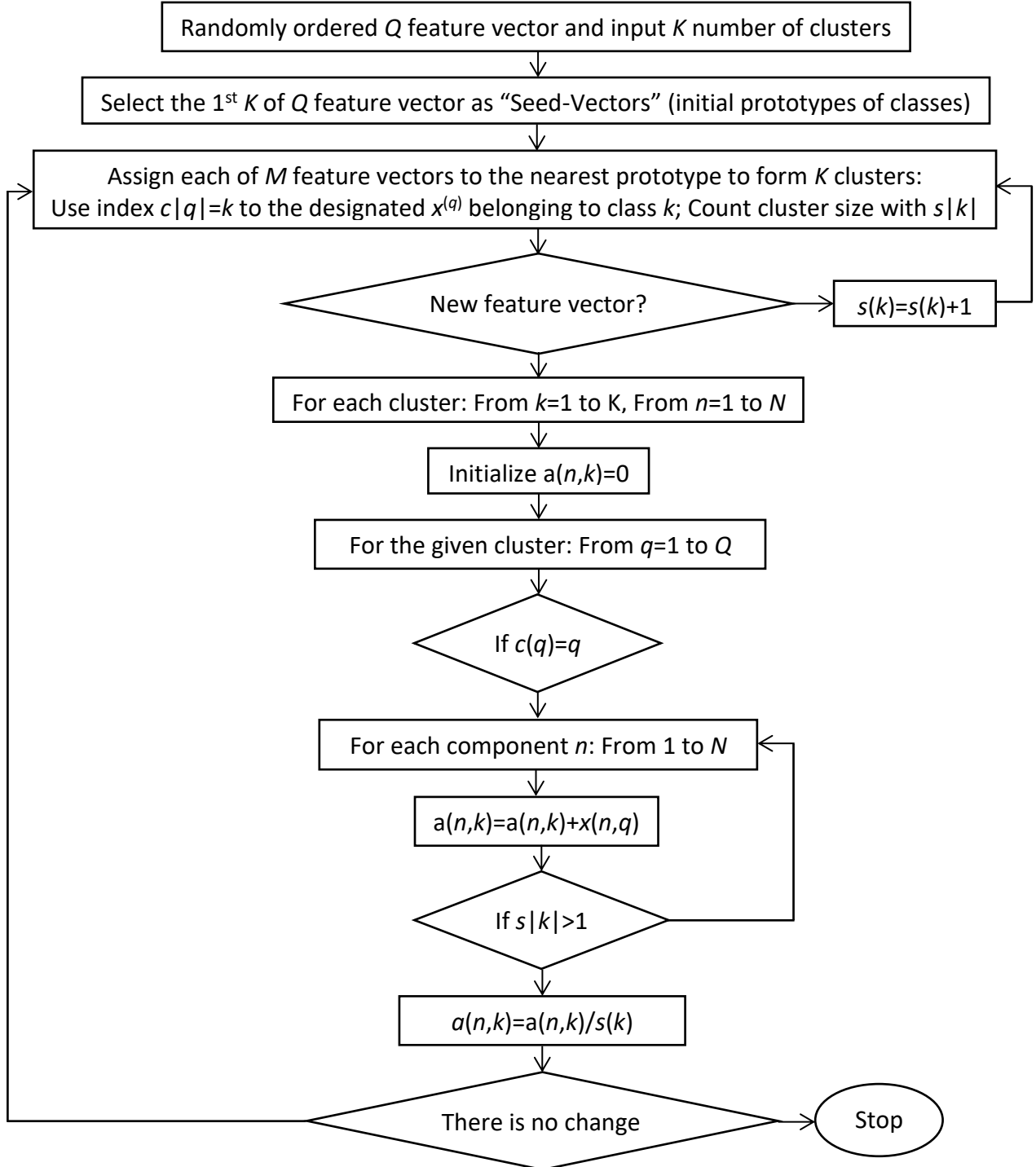
- Open source MATLAB code for PNCC by Chanwoo Kim, Carnegie Mellon University.
- Codes of MATLAB audio processing by Dan Ellis, Columbia University.
- MATLAB code of Hidden Markov Model Toolkit (HTK), Cambridge University.

- The MATLAB VOICEBOX toolbox for speech processing, Imperial Collage-London.
- Machine Learning Toolbox (MLT) by Roger Jang.
- Speaker diarization tool-kit, GitHub Institute.
- MATLAB toolbox for performance measurement in source separation (BSS Eval).

The MS-OFFICE applications are utilized to tabulate the database *FileName.xls* files and to edit the ASCII-code files. For speech-DSP, the open source application AUDACITY is utilized to edit, display, play and compare between different *FileName.wav* speech files. For the rich-text editing, NOTEPAD++ is used.

Appendix C. k-means Flowchart

There are Q data: d_1, d_2, \dots, d_Q . To cluster those data into k clusters, Lloyd standard k-means algorithm [139] can achieve that using the following flowchart:



Appendix D. NMF Procedure

There is $[S]$ matrix with the dimension $r \times c$. To Factorize $[S]=[W] \times [H]$ (where $[W]$ is $r \times ss$ and $[H]$ is $ss \times c$), Lee algorithm is the following sequence [74, 75]:

Initialize $[W(0)] \geq \alpha$ and $[H(0)] \geq \alpha$ randomly;

$k \leftarrow 0$;

Repeat:

$$[D(k)] = [[H(k)] \times [H^T(k)]];$$

$$[Q(k)] = [V] \times [H^T(k)];$$

for $i = 1:ss$

$$E_{rc}(k) = [D(k)]_{rc} (1 - \delta_{ri} \delta_{ic});$$

$\forall r$;

$c \in \{1, 2, \dots, ss\}$;

$$[W_i(k+1)] \leftarrow \max\left\{\frac{[Q(k)]_i - [W_i(k+1)], \dots, [W_{i-1}(k+1)], [W_i(k)], \dots, [W_{ss}(k)] \times [E_i(k)]}{[D(k)]_{ii}}, \alpha\right\};$$

end

$$[C(k)] = [[W^T(k+1)] \times [W(k+1)]];$$

$$[R(k)] = [W^T(k+1)] \times [V];$$

For $j = 1:ss$

$$F_{rc}(k) = [C(k)]_{rc} (1 - \delta_{rj} \delta_{jc});$$

$\forall r$;

$c \in \{1, 2, \dots, ss\}$;

$$[H_j(k+1)] \leftarrow \max\left\{\frac{[R(k)]_j - [F_j(k)] \times [H_1^T(k+1)], \dots, [H_{j-1}^T(k+1)], [H_j^T(k)], \dots, H_{ss}^T(k)]}{[C(k)]_{jj}}, \alpha\right\};$$

end

Until:

$$\frac{\| [V] - [W(k)] \times [H(k)] \|_F - \| [V] - [W(k+1)] \times [H(k+1)] \|_F}{\| [V] \|_F} < \varepsilon$$

% ε is the stopping threshold, relatively is very small value e.g. 10^{-5}

References

- [1] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*: Prentice Hall, 1978.
- [2] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete-time processing of speech signals*: Wiley-IEEE Press, 1999.
- [3] H. Beigi, *Fundamentals of speaker recognition*: Springer Science & Business Media, 2011.
- [4] N. Morgan, H. Bourland, S. Greenberg, H. Hermansky, and W. Su-Lin, "Stochastic perceptual models of speech," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, 1995, pp. 397-400 vol.1.
- [5] O. Ghitza and M. M. Sondhi, "Hidden Markov models with templates as non-stationary states: an application to speech recognition," *Computer Speech & Language*, vol. 7, pp. 101-120, 1993.
- [6] E. Sejdić, I. Djurović, and J. Jiang, "Time--frequency feature representation using energy concentration: An overview of recent advances," *Digital Signal Processing*, vol. 19, pp. 153-183, 2009.
- [7] A. M. Noll, "Cepstrum pitch determination," *The journal of the acoustical society of America*, vol. 41, pp. 293-309, 1967.
- [8] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.
- [9] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4353-4356.
- [10] D. Charlet, C. Barras, and J.-S. Liénard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7707-7711.
- [11] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 356-370, 2012.
- [12] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," *Idiap2013*.
- [13] H. A. Kadhim, L. Woo, and S. Dlay, "Speaker diarization by dependent combination of audio features," *International Journal of Simulation--Systems, Science & Technology IJtauT*, 16(1), 2016.
- [14] H. A. Kadhim, L. Woo, and S. Dlay, "Statistical Speaker Diarization Using Dependent Combination of Extracted Features," in *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, 2015, pp. 291-296.
- [15] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, pp. 621-633, 2013.
- [16] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1381-1390, 2013.
- [17] K. Fukunaga, *Introduction to statistical pattern recognition*: Academic press, 2013.
- [18] E. Alpaydin, *Introduction to machine learning*: MIT press, 2014.

- [19] V. Vapnik, *The nature of statistical learning theory*: Springer Science & Business Media, 2013.
- [20] S. Fernández, A. Graves, and J. Schmidhuber, "Phoneme recognition in TIMIT with BLSTM-CTC," *arXiv preprint arXiv:0804.3269*, 2008.
- [21] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013, pp. 1-4.
- [22] H. A. Kadhimi, L. Woo, and S. Dlay, "Overlapped-speech detection and blind speech separation of spontaneous conversation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, Submitted, 2017.
- [23] H. A. Kadhimi, L. Woo, and S. Dlay, "Overlapped-speech detection and informed speech separation of spontaneous conversation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, Submitted, 2017.
- [24] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2046-2057, 2011.
- [25] P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin, "Subjective and objective quality assessment of single-channel speech separation algorithms," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 69-72.
- [26] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, pp. 2578-2583, 2008.
- [27] X. Anguera, C. Wooters, and J. Hernando, "Purity algorithms for speaker diarization of meetings data," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006, pp. I-I.
- [28] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, *et al.*, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928-1936, 2012.
- [29] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1766-1776, 2007.
- [30] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined Convolutional Blind Source Separation via Time-Frequency Masking," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 18, pp. 101-116, 2010.
- [31] O. Ghizta, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech & Language*, vol. 1, pp. 109-130, 1986.
- [32] O. Ghizta, "Auditory nerve representation criteria for speech analysis/Synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, pp. 736-740, 1987.
- [33] S. Furui, "Toward robust speech recognition under adverse conditions," in *Speech Processing in Adverse Conditions*, 1992.
- [34] D. Kobayashi, S. Kajita, K. Takeda, and F. Itakura, "Extracting speech features from human speech like noise," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, pp. 418-421 vol.1.
- [35] T. Taniguchi, S. Kajita, K. Takeda, and F. Itakura, "Blind signal separation for recognizing overlapped speech," *Journal of the Acoustical Society of Japan (E)*, vol. 19, pp. 385-390,

- 1998.
- [36] X. A. Miro, *Robust speaker diarization for meetings*: Universitat Politècnica de Catalunya, 2007.
 - [37] K. A. Boakye, *Audio segmentation for meetings speech processing*: ProQuest, 2008.
 - [38] S. Otterson, "Use of speaker location features in meeting diarization," University of Washington, 2008.
 - [39] E. K. C. Wei, "Speaker Diarization of News Broadcasts and Meeting Recordings," PhD, Nanyang Technological University, 2008.
 - [40] B. Trueba-Hornero, "Handling overlapped speech in speaker diarization," *Master's thesis, Universitat Politècnica de Catalunya*, 2008.
 - [41] S. Shum, "Unsupervised methods for speaker diarization," Massachusetts Institute of Technology, 2011.
 - [42] D. Wang, "Speaker Diarization "Who Spoke When"," PhD, Queensland University of Technology, 2012.
 - [43] J. L. Serrano, "Speaker Diarization and Tracking in Multiple-Sensor Environments," PhD thesis, Universitat Politècnica de Catalunya, 2012.
 - [44] M. T. Knox, "Speaker Diarization: Current Limitations and New Directions," PhD, University of California-Berkeley, 2013.
 - [45] N. T. Hieu, "Speaker diarization in meetings domain," PhD, Nanyang Technological University, 2014.
 - [46] S. Bozonnet, "New insights into hierarchical clustering and linguistic normalization for speaker diarization," Télécom ParisTech, 2012.
 - [47] H. D. Flores, "Fast Cross-Session Speaker Diarization," PhD, Universitat Autònoma de Barcelona, 2015.
 - [48] S. H. Yella, "Speaker diarization of spontaneous meeting room Conversations," PhD, GÉNIE ÉLECTRIQUE ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, 2015.
 - [49] S. Wegmann, P. Zhan, I. Carp, M. Newman, J. Yamron, and L. Gillick, "Dragon systems' 1998 broadcast news transcription system," in *Proc. 1999 DARPA Broadcast News Workshop*, 1999, pp. 277-280.
 - [50] X. Guo, W. Zhu, Q. Shi, S. Chen, and R. Gopinath, "The IBM LVCSR system used for 1998 Mandarin broadcast news transcription evaluation," in *Proc. DARPA Broadcast News Workshop*, 1999, pp. 179-182.
 - [51] A. Oak, "It's a Nice Idea, but it's not actually Real: Assessing the Objects and Activities of Design," *Journal of Art & Design Education*, vol. 19, pp. 86-95, 2000.
 - [52] L. Grönqvist, "The MultiTool User's Manual," *A tool for browsing and synchronizing transcribed dialogues and corresponding video recordings. Göteborg University, Department of Linguistics*, 2000.
 - [53] T. Giannakopoulos, "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis," *PloS one*, vol. 10, p. e0144610, 2015.
 - [54] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, pp. 5-22, 2001.
 - [55] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: improving speaker diarization by automatically identifying and excluding overlapped speech," in *INTERSPEECH*, 2008, pp.

- 32-35.
- [56] K. Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
 - [57] L.-R. Dai, Y. Song, and R.-H. Wang, "A Study on the Frame Synchronized Segregation of Overlapped Speech," *Acta Electronica Sinica*, vol. 30, pp. 1552-1554, 2002.
 - [58] K. Laskowski and T. Schultz, "Unsupervised Learning of Overlapped Speech Model Parameters For Multichannel Speech Activity Detection in Meetings," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006, pp. I-I.
 - [59] O. Ben-Harush, H. Guterman, and I. Lapidot, "Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization," in *2009 IEEE International Workshop on Machine Learning for Signal Processing*, 2009, pp. 1-6.
 - [60] V. Rozgic, K. J. Han, P. G. Georgiou, and S. Narayanan, "Multimodal Speaker Segmentation in Presence of Overlapped Speech Segments," in *2008 Tenth IEEE International Symposium on Multimedia*, 2008, pp. 679-684.
 - [61] B. Xiao, P. K. Ghosh, P. Georgiou, and S. S. Narayanan, "Overlapped speech detection using long-term spectro-temporal similarity in stereo recording," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5216-5219.
 - [62] H. Pericás and F. Javier, "On the improvement of speaker diarization by detecting overlapped speech," in *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*, 2010, pp. 153-156.
 - [63] R. Yokoyama, Y. Nasu, K. Shinoda, and K. Iwano, "Overlapped Speech Detection in Meeting Using Cross-Channel Spectral Subtraction and Spectrum Similarity," *InterSpeech2012*, 2012.
 - [64] L. Wei, H. Qian-Hua, L. Yan-Xiong, Z. Xue-Yuan, and F. Xiao-Hvi, "Fractal dimension feature for distinguishing between overlapped speech and single-speaker speech," in *2012 International Conference on Machine Learning and Cybernetics*, 2012, pp. 148-151.
 - [65] J. Málek, Z. Koldovský, and P. Tichavský, "Semi-blind source separation based on ICA and overlapped speech detection," in *International Conference on Latent Variable Analysis and Signal Separation*, 2012, pp. 462-469.
 - [66] A. Z. Wang, C. G. Bi, and B. X. Li, "Blind separation method of overlapped speech mixtures in STFT domain with noise and residual crosstalk suppression," in *2016 12th IEEE International Conference on Control and Automation (ICCA)*, 2016, pp. 876-880.
 - [67] S. Shum, N. Dehak, E. Chuangsuwanich, D. A. Reynolds, and J. R. Glass, "Exploiting Intra-Conversation Variability for Speaker Diarization," in *INTERSPEECH*, 2011, pp. 945-948.
 - [68] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 1059-1070, 2010.
 - [69] M.-J. Caraty and C. Montacié, "Detecting Speech Interruptions for Automatic Conflict Detection," in *Conflict and Multimodal Communication*, ed: Springer, 2015, pp. 377-401.
 - [70] H. A. Kadhim, L. Woo, and S. Dlay, "Stochastic Overlapped-Speech Detection of Spontaneous Conversation," *Submitted; IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 2017.
 - [71] H. A. Kadhim, L. Woo, and S. Dlay, "Novel algorithm for speech segregation by optimized

- k-means of statistical properties of clustered features," in *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)*, 2015, pp. 286-291.
- [72] H. A. Kadhim, L. Woo, and S. Dlay, "Statistical speech Segregation using the developed k-means of audio feature," presented at the IEEE International Conference on Image Information Processing (ICIIP -2015), 2015.
 - [73] W. H. Lawton and E. A. Sylvestre, "Self modeling curve resolution," *Technometrics*, vol. 13, pp. 617-633, 1971.
 - [74] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556-562.
 - [75] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.
 - [76] P. O. Hoyer, "Non-negative sparse coding," in *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 557-565.
 - [77] J. Eggert and E. Korner, "Sparse coding and NMF," in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, 2004, pp. 2529-2533 vol.4.
 - [78] M. N. Schmidt and R. K. Olsson, "Linear Regression on Sparse Features for Single-Channel Speech Separation," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 26-29.
 - [79] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Spoken Language Processing, ISCA International Conference on (INTERSPEECH)*, 2006.
 - [80] X. Downie, C. Laurier, and M. Ehmann, "The 2007 MIREX audio mood classification task: Lessons learned," in *Proc. 9th Int. Conf. Music Inf. Retrieval*, 2008, pp. 462-467.
 - [81] E. Vincent, N. Bertin, and R. Badeau, "Two nonnegative matrix factorization methods for polyphonic pitch transcription," in *2007 Music Information Retrieval Evaluation eXchange (MIREX)*, 2007.
 - [82] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, pp. 793-830, 2009.
 - [83] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 550-563, 2010.
 - [84] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 2012, pp. 1-6.
 - [85] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 1825-1828.
 - [86] W. Wang, "Instantaneous vs. Convolutional Non-Negative Matrix Factorization: Models, Algorithms and Applications," *Machine Audition: Principles, Algorithms and Systems: Principles, Algorithms and Systems*, p. 353, 2010.
 - [87] H. A. Kadhim, L. Woo, and S. Dlay, "Speech separation of spontaneous conversation by filter-bank, NMF and speaker clustering," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, Submitted, 2017.

- [88] B. Cauchi, "Non-negative matrix factorisation applied to auditory scenes classification," *Master's thesis, Master ATIAM, Université Pierre et Marie Curie*, 2011.
- [89] N. Zheng, Y. Cai, X. Li, and T. Lee, "Classifying NMF components based on vector similarity for speech and music separation," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012, pp. 1-6.
- [90] N. Moritz, M. R. Schädler, K. Adiloglu, B. T. Meyer, T. Jürgens, T. Gerkmann, *et al.*, "Noise robust distant automatic speech recognition utilizing NMF based source separation and auditory feature extraction," *Proc. of CHiME*, pp. 1-6, 2013.
- [91] J. R, "Non-negative matrix factorization based algorithms to cluster frequency basis functions for monaural sound source separation," PhD, Dublin Institute of Technology, 2013.
- [92] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised single-channel separation of nonstationary signals using Gammatone filterbank and Itakura–Saito nonnegative matrix two-dimensional factorizations," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, pp. 662-675, 2013.
- [93] B. Gao, W. L. Woo, and S. S. Dlay, "Variational regularized 2-D nonnegative matrix factorization," *IEEE transactions on neural networks and learning systems*, vol. 23, pp. 703-716, 2012.
- [94] B. Gao, W. L. Woo, and S. S. Dlay, "Adaptive sparsity non-negative matrix factorization for single-channel source separation," *IEEE journal of selected topics in signal processing*, vol. 5, pp. 989-1001, 2011.
- [95] Q. Liu, W. Wang, P. J. Jackson, M. Barnard, J. Kittler, and J. Chambers, "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 61, pp. 5520-5535, 2013.
- [96] Q. Liu, S. M. Naqvi, W. Wang, P. Jackson, and J. Chambers, "Robust feature selection for scaling ambiguity reduction in audio-visual convolutive BSS," in *Signal Processing Conference, 2011 19th European*, 2011, pp. 1060-1064.
- [97] A. Kazemi, R. Boostani, and F. Sobhanmanesh, "Audio visual speech source separation via improved context dependent association model," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, pp. 1-16, 2014.
- [98] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1699-1712, 2013.
- [99] L. Zhen, D. Peng, Z. Yi, Y. Xiang, and P. Chen, "Underdetermined Blind Source Separation Using Sparse Coding," *IEEE Transactions on Neural Networks and Learning Systems*, 2016.
- [100] W. Nogueira, T. Gajecski, B. Krueger, J. Janer, and A. Buechner, "Development of a sound coding strategy based on a deep recurrent neural network for monaural source separation in cochlear implants," in *Speech Communication; 12. ITG Symposium; Proceedings of*, 2016, pp. 1-5.
- [101] S. Arberet and P. Vandergheynst, "Reverberant audio source separation via sparse and low-rank modeling," *IEEE Signal Processing Letters*, vol. 21, pp. 404-408, 2014.
- [102] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation with probabilistic reverberation modeling," in *Applications of Signal Processing to Audio and*

- Acoustics (WASPAA), 2015 IEEE Workshop on*, 2015, pp. 1-5.
- [103] A. Asaei, M. Golbabaei, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 620-633, 2014.
 - [104] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, pp. 116-124, 2014.
 - [105] J. Driedger, H. Grohgan, T. Prätzlich, S. Ewert, and M. Müller, "Score-informed audio decomposition and applications," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 541-544.
 - [106] Z.-Q. Wang, Y. Zhao, and D. Wang, "Phoneme-specific speech separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 146-150.
 - [107] R. Hennequin, J. J. Burred, S. Maller, and P. Leveau, "Speech-guided source separation using a pitch-adaptive guide signal model," in *ICASSP*, 2014, pp. 6672-6676.
 - [108] Q. Wang, W. Woo, and S. Dlay, "Informed Single-Channel Speech Separation Using HMM-GMM User-Generated Exemplar Source," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 2087-2100, 2014.
 - [109] A. Lefevre, "Dictionary learning methods for single-channel source separation," École normale supérieure de Cachan-ENS Cachan, 2012.
 - [110] M. Fakhry, P. Svaizer, and M. Omologo, "Reverberant audio source separation using partially pre-trained nonnegative matrix factorization," in *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*, 2014, pp. 273-277.
 - [111] T. Ming, X. Xiang, and J. Yishan, "Nmf based speech and music separation in monaural speech recordings with sparseness and temporal continuity constraints," in *3rd International Conference on Multimedia Technology (ICMT-13)*, 2013.
 - [112] C. Joder, F. Weninger, D. Virette, and B. Schuller, "A comparative study on sparsity penalties for NMF-based speech separation: Beyond LP-norms," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 858-862.
 - [113] J. L. Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 66-70.
 - [114] S. Nie, S. Liang, H. Li, X. Zhang, Z. Yang, W. J. Liu, *et al.*, "Exploiting spectro-temporal structures using NMF for DNN-based supervised speech separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 469-473.
 - [115] D. Bouvier, N. Obin, M. Liuni, and A. Roebel, "A source/filter model with adaptive constraints for NMF-based speech separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 131-135.
 - [116] B. Gao, W. Woo, and S. Dlay, "Single-channel source separation using EMD-subband variable regularized sparse features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 961-976, 2011.
 - [117] H. A. Kadhim, L. Woo, and S. Dlay, "Informed speech separation of spontaneous conversation by semi-supervised NMF," *IEEE/ACM Transactions on Audio, Speech and*

Language Processing (TASLP), Submitted, 2017.

- [118] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 888-891.
- [119] F. J. Rodriguez-Serrano, S. Ewert, P. Vera-Candeas, and M. Sandler, "A score-informed shift-invariant extension of complex matrix factorization for improving the separation of overlapped partials in music recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 61-65.
- [120] N. Souviraà-Labastie, E. Vincent, and F. Bimbot, "Music separation guided by cover tracks: designing the joint NMF model," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 484-488.
- [121] N. Guan, L. Lan, D. Tao, Z. Luo, and X. Yang, "Transductive nonnegative matrix factorization for semi-supervised high-performance speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 2534-2538.
- [122] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*, ed: Springer, 2008, pp. 509-519.
- [123] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1557-1565, 2006.
- [124] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation* vol. 615: Springer, 2007.
- [125] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," *Multichannel Speech Processing Handbook*, pp. 1065-1084, 2007.
- [126] D. O'shaughnessy, *Speech communication: human and machine*: Universities press, 1987.
- [127] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1315-1329, 2016.
- [128] P. Cano, E. Batle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting," in *2002 IEEE Workshop on Multimedia Signal Processing.*, 2002, pp. 169-173.
- [129] J. Bradbury, "Linear predictive coding," *Mc G. Hill*, 2000.
- [130] D. O. Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, pp. 29-32, 1988.
- [131] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, pp. 1738-1752, 1990.
- [132] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *The Bell System Technical Journal*, vol. 12, pp. 377-430, 1933.
- [133] O. Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *The Journal of the Acoustical Society of America*, vol. 110, pp. 1628-1640, 2001.
- [134] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, pp. 103-138, 1990.
- [135] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, pp. 248-248, 1961.
- [136] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, 1991, pp. 121-124.

- [137] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, pp. 578-589, 1994.
- [138] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, pp. 264-323, 1999.
- [139] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, pp. 129-137, 1982.
- [140] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027-1035.
- [141] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," *Proceedings of the VLDB Endowment*, vol. 5, pp. 622-633, 2012.
- [142] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.
- [143] S. Bozonnet, N. Evans, C. Fredouille, D. Wang, and R. Troncy, "An integrated top-down/bottom-up approach to speaker diarization," in *Interspeech 2010, September 26-30, Makuhari, Japan*, 2010, pp. Interspeech 2010, September 26-30, Makuhari, Japan.
- [144] G. Karypis, E.-H. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, pp. 68-75, 1999.
- [145] N. Evans, S. Bozonnet, D. Wang, C. Fredouille, and R. Troncy, "A comparative study of bottom-up and top-down approaches to speaker diarization," *IEEE Transactions on Audio, speech, and language processing*, vol. 20, pp. 382-392, 2012.
- [146] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127-132.
- [147] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system," in *RT-04F Workshop*, 2004, p. 23.
- [148] P. Smaragdis, "Convolutional Speech Bases and Their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1-12, 2007.
- [149] E. Oja, "The nonlinear PCA learning rule in independent component analysis," *Neurocomputing*, vol. 17, pp. 25-45, 1997.
- [150] P. Comon, "Independent component analysis, a new concept?," *Signal processing*, vol. 36, pp. 287-314, 1994.
- [151] D. P. Ellis, "Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures," *Speech Communication*, vol. 27, pp. 281-298, 1999.
- [152] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, pp. 1830-1847, 2004.
- [153] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the national academy of sciences*, vol. 101, pp. 4164-4169, 2004.
- [154] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, 2009.
- [155] A. Janeczek and Y. Tan, "Iterative improvement of the multiplicative update nmf algorithm

- using nature-inspired optimization," in *Natural Computation (ICNC), 2011 Seventh International Conference on*, 2011, pp. 1668-1672.
- [156] S. Yang and M. Ye, "Global minima analysis of Lee and Seung's NMF algorithms," *Neural processing letters*, vol. 38, pp. 29-51, 2013.
 - [157] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, 2003, pp. 411-416.
 - [158] H. Jin, F. Kubala, and R. Schwartz, "Automatic speaker clustering," in *Proceedings of the DARPA speech recognition workshop*, 1997, pp. 108-111.
 - [159] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," *Signal processing*, vol. 88, pp. 1091-1124, 2008.
 - [160] Q.-H. Lin and Y.-G. Hao, "A survey of semi-blind ICA for speech separation in frequency domain," in *Green Circuits and Systems (ICGCS), 2010 International Conference on*, 2010, pp. 632-636.
 - [161] X. Anguera, "Speaker diarization: A review of recent research (vol 20, pg 356, 2012)," *IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING*, vol. 21, pp. 1308-1308, 2013.
 - [162] e. a. J. Garofolo. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Available (in July/2017): <https://catalog.ldc.upenn.edu/LDC93S1>
 - [163] S. Nie, S. Liang, H. Li, X. Zhang, Z. Yang, W. J. Liu, *et al.*, "Exploiting spectro-temporal structures using NMF for DNN-based supervised speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 469-473.
 - [164] J. Janer and R. Marxer, "Separation of unvoiced fricatives in singing voice mixtures with semi-supervised NMF," in *Proc. 16th Int. Conf. Digital Audio Effects*, 2013, pp. 2-5.
 - [165] S. J. Chapman, *Essentials of MATLAB programming*: Cengage Learning, 2016.
 - [166] P. Magron, R. Badeau, and B. David, "Phase recovery in NMF for audio source separation: an insightful benchmark," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 81-85.
 - [167] W. Han, X. Zhang, J. Yang, M. Sun, and G. Min, "Joint Optimization of a Perceptual Modified Wiener Filtering Mask and Deep Neural Networks for Monaural Speech Separation," in *Pacific Rim Conference on Multimedia*, 2016, pp. 469-478.
 - [168] S. U. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind Speech Separation and Enhancement with GCC-NMF," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.
 - [169] Y. Salaün, E. Vincent, N. Bertin, N. Souvira-Labastie, X. Jaureguiberry, D. T. Tran, *et al.*, "The flexible audio source separation toolbox version 2.0," in *ICASSP*, 2014.
 - [170] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1118-1133, 2012.
 - [171] K. Adiloğlu and E. Vincent, "Variational Bayesian inference for source separation and robust feature extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1746-1758, 2016.
 - [172] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *International Conference on Latent*

- Variable Analysis and Signal Separation*, 2012, pp. 322-329.
- [173] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*, 2015, pp. 387-395.
 - [174] Z. Rafii and B. Pardo, "Online REPET-SIM for real-time speech enhancement," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 848-852.
 - [175] S. Arberet, A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, *et al.*, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*, 2010, pp. 1-4.
 - [176] L. Le Magoarou, A. Ozerov, and N. Q. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, 2013, pp. 1-6.
 - [177] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE transactions on audio, speech, and language processing*, vol. 19, pp. 549-557, 2011.
 - [178] E. Gopi, *Digital speech processing using Matlab*: Springer, 2014.